On the Finite-Time Complexity and Practical Computation of Approximate Stationarity Concepts of Lipschitz Functions

Lai Tian [1] **Kaiwen Zhou** [2] **Anthony Man-Cho So** [1]

Abstract

We report a practical finite-time algorithmic scheme to compute approximately stationary points for nonconvex nonsmooth Lipschitz functions. In particular, we are interested in two kinds of approximate stationarity notions for nonconvex nonsmooth problems, i.e., Goldstein approximate stationarity (GAS) and near-approximate stationarity (NAS). For GAS, our scheme removes the unrealistic subgradient selection oracle assumption in (Zhang et al., 2020, Assumption 1) and computes GAS with the same finite-time complexity. For NAS, Davis & Drusvyatskiy (2019) showed that ρ-weakly convex functions admit finite-time computation, while Tian & So (2021) provided the matching impossibility results of dimension-free finite-time complexity for first-order methods. Complement to these developments, in this paper, we isolate a new class of functions that could be Clarke irregular (and thus not weakly convex anymore) and show that our new algorithmic scheme can compute NAS points for functions in that class within finite time. To demonstrate the wide applicability of our new theoretical framework, we show that ρ-margin SVM, 1-layer, and 2-layer ReLU neural networks, all being Clarke irregular, satisfy our new conditions.

1. Introduction

In this paper, we consider the following general optimization problem for an L-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$

$$\min_{x\in\mathbb{R}^d} f(x), \qquad (\Diamond)$$

where f could be both nonsmooth and nonconvex ("non"-setting for short). We are particularly interested in algorithms with a finite-time complexity for computing approximately stationary points of Problem ($\Diamond$). Note that when f is smooth, it is folkloric that computing an ϵ-stationary point (i.e., $\|\nabla f(x)\| \le \epsilon$) only requires $O(\epsilon^{-2})$ calls, which is dimension-independent and finite, to the gradient oracle with gradient descent (Nemirovskij & Yudin, 1983).

In the general Lipschitz "non"-setting, a widely used generalized subdifferential $\partial f(x)$ is due to Clarke (1990, Section 2.1) (see also Definition 2.1), which reduces to the convex subdifferential (resp. gradient) if f is convex (resp. smooth). Therefore, by mimicking results in the smooth scenario, it is natural to conjecture that we may be able to design algorithms to compute elements in $\{x : \mathrm{dist}(0, \partial f(x)) \le \epsilon\}$ in finite time with high probability. However, as shown by Zhang et al. (2020, Theorem 5), that is impossible for any first-order method. Thus, it is curious to ask: What kind of approximate stationarity concept in the "non"-setting will admit dimension-free finite-time computation?

Davis & Drusvyatskiy (2019) gave a nice answer for the class of ρ-weakly convex functions[1] by introducing a notion named near-approximate stationarity (NAS, see Definition 2.5), which is closely related to the gradient of the Moreau envelope of f. They showed that a subgradient-type method computes an (ϵ, δ)-NAS point with $O(\rho^4\delta^{-4} + \epsilon^{-4})$ calls to the subgradient oracle. However, many modern ML models are indeed not weakly convex,[2] e.g., neural networks with ReLU activation functions. Even worse, by extending the Lipschitz hardness results in (Kornowski & Shamir, 2021), Tian & So (2021) demonstrated that, for any finite T, there exists a finte $\rho(T)$ such that, for any $0 \le \epsilon, \delta < \frac{1}{2}$ uniformly, computing an (ϵ, δ)-NAS point for $\rho(T)$-weakly convex functions within T steps is impossible.

On the other front, starting from the seminal work of Goldstein (1977), a notion named Goldstein approximate stationarity (GAS, see Definition 2.1) exhibits favorable algorithmic consequences. The story begins with an approximation

[1]Department of Systems Engineering & Engineering Management, The Chinese University of Hong Kong, Shatin, NT, Hong Kong [2]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, NT, Hong Kong. Correspondence to: Anthony Man-Cho So <manchoso@se.cuhk.edu.hk>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

[1]Recall f is ρ-weakly convex if $f(x) + \frac{\rho}{2}\|x\|^2$ is convex. Weak convexity implies Clarke regularity (Vial, 1983, Proposition 4.5).

[2]$g(x) = -\max\{x, 0\}$ is not Clarke regular (cf. (Clarke, 1990, Definition 2.3.4)) and not ρ-weakly convex for any $\rho \in \mathbb{R}$.

of the Clarke subdifferential $\partial_\delta f(x)$ (see Definition 2.2). If we update iteratively with

$$x_{k+1} \leftarrow x_k - \delta \cdot g_k/\|g_k\|,$$

where $g_k := \arg\min_{g\in\partial_\delta f(x_k)} \|g\|$ is the minimal norm element in $\partial_\delta f(x_k)$, then we can compute an (ϵ,δ)-GAS point in $O(\epsilon^{-1}\delta^{-1})$ steps. The problem is that obtaining g_k for a general Lipschitz function can be computationally expensive (if possible at all) as there is no known approach to compute $\partial_\delta f(x)$. Therefore, a series of works, e.g., (Burke et al., 2020), proposed to build a polyhedral approximation of $\partial_\delta f(x_k)$ via random sampling and compute an approximate g_k by solving a QP in every iteration. However, the number of sampling points needed for meaningful approximation of $\partial_\delta f(x_k) \subseteq \mathbb{R}^d$ is lower bounded by the dimension d. Thus, a dimension-free finite-time complexity cannot be achieved within the existing gradient sampling scheme.

Recently, Zhang et al. (2020) introduced a novel algorithm that computes (ϵ,δ)-GAS points for general Lipschitz functions with a dimension-free finite-time complexity $\widetilde{O}(\epsilon^{-3}\delta^{-1})$. They provided a randomized procedure that compute $\tilde{g}_k : \langle \tilde{g}_k, \partial_\delta f(x_k)\rangle \le \frac{1}{4}\|\tilde{g}_k\|^2$ with high probability within $\widetilde{O}(\epsilon^{-2})$ oracle calls. However, their assumption on the subgradient oracle is stringent and hard to be implemented in practice. We restate their assumption below.

Oracle in (Zhang et al., 2020, Assumption 1(a)). Given x, d, the oracle $\mathbb{O}(x,d)$ returns $f(x)$ and a Clarke subgradient g_x, such that $g_x \in \partial f(x)$ satisfies $\langle g_x, d\rangle = f'(x;d)$.

Indeed, even computing an *arbitrary* element in $\partial f(x)$ for general Lipschitz functions is highly nontrivial (Burke et al., 2002; Nesterov, 2005; Khan & Barton, 2013; Kakade & Lee, 2018), let alone the required subgradient needs to satisfy certain linear equation, which was recognized very early on as impractical (Wolfe, 1975). Such considerations motivate the following question (**Q1**):

Can we compute GAS points in finite time with a provable and practical algorithm?

It is notable that GAS is a strictly weaker[3] stationarity notion than NAS even for continuously differentiable functions (Kornowski & Shamir, 2021, Proposition 1) and convex functions (Proposition 2.7). However, the computability of NAS is much worse than that of GAS, since finite-time algorithms for NAS only exist if the objective function is ρ-weakly convex, which rules out many interesting machine learning models. Thus, it is of interest to ask (**Q2**):

Can we compute NAS points in finite time for functions beyond ρ-weakly convex, practically?

[3]Formally, if x is (ϵ,δ)-GAS, then x is also (ϵ,δ)-NAS. See Definitions 2.4 and 2.5.

1.1. Prior Arts

Asymptotic Analysis. The asymptotic computability of Clarke stationary points (i.e., $\{x : 0 \in \partial f(x)\}$) has been well-understood for quite general functions. With a differential inclusion perspective, Benaïm et al. (2005); Majewski et al. (2018); Davis et al. (2020) studied the asymptotic convergence of subgradient-type methods. In particular, Davis et al. (2020) proved the asymptotic convergence to Clarke stationary points for Whitney stratifiable objective functions, which include deep ReLU neural networks as a special case. Daniilidis & Drusvyatskiy (2020) demonstrated that the vanilla subgradient method may not converge for general Lipschitz functions even in continuous time.

Finite-Time Analysis. In contrast to the asymptotic regime, the finite-time complexity in the general "non"-setting is still developing. On the positive side, (Davis & Grimmer, 2019; Davis & Drusvyatskiy, 2019) showed that for ρ-weakly convex functions, (ϵ,δ)-NAS is computable with $O(\rho^4\delta^{-4}+\epsilon^{-4})$ oracle calls. On the negative side, Kornowski & Shamir (2021) showed that computing NAS for Lipschitz functions in dimension-independent finite time is impossible. Tian & So (2021) sharpened the hardness results for NAS to ρ-weakly convex with unbounded ρ, thus matching the positive results. For GAS, the gradient sampling scheme (Burke et al., 2005; Kiwiel, 2007; 2010; Burke et al., 2020) promises finite but dimension-dependent complexity. (Zhang et al., 2020) reported a novel dimension-independent finite-time algorithm with a impractical subgradient oracle. A recent concurrent work (Davis et al., 2021) adopted similar strategy as our Section 3.2 with different algorithmic implementation. Another line of research is to exploit structure: Duchi & Ruan (2018); Drusvyatskiy & Paquette (2019); Davis & Drusvyatskiy (2019); Bolte et al. (2018); Beck & Hallak (2020). In these settings, nonsmoothness and nonconvexity are properly separated making finite-time analysis possible.

1.2. Contributions

We highlight the main contributions as follows.

- For **Q1**, we report a practical algorithmic scheme to compute GAS points for general Lipschitz functions with finite-time complexity in both deterministic and stochastic settings.
- For **Q2**, we isolate a new function class within which our new algorithmic scheme computes NAS points in finite time.This goes far beyond existing ρ-weakly convex results. Besides, we establish a series of theoretical tools to compute parameters in our new function class.
- To demonstrate the wide applicability of the new theoretical framework, we show that ρ-margin SVM, 1-layer, and 2-layer ReLU neural networks, all being Clarke irregular, satisfy our new conditions.

Notations. The notation used in this paper is mostly standard in variational analysis. $\mathrm{dist}(x,S) :- \inf_{v\in S}\|v-x\|$; $A\oplus B$ denotes the direct sum of A and B; $\mathbb{B}_\epsilon(x) :- \{v : \|v-x\|\le\epsilon\}$; $\mathbb{B} := \mathbb{B}_1(0)$; $\mathrm{Co}\,S$ is the convex hull of set S; $\pi_1 A := \{x : \exists y, (x,y)\in A\}$; A^c is the complement of set A; $\mathrm{Vec}(X)$ is the vectorization of matrix X; $\mathbb{S}^{d-1} := \{x\in\mathbb{R}^d : \|x\|=1\}$.

2. Preliminaries

In this section, we introduce the necessary background on variational analysis for Lipschitz functions. To begin, we recall the following definition of Clarke subdifferential (Rockafellar & Wets, 2009, Theorem 9.61).

Definition 2.1 (Clarke subdifferential). *Given a point x, the Clarke subdifferential of Lipschitz f at x is defined by*

$$\partial f(x) := \mathrm{Co}\left\{s : \exists x'\to x, \nabla f(x') \text{ exists}, \nabla f(x')\to s\right\}.$$

The following δ-approximation of Clarke subdifferential introduced by Goldstein (1977) has nice theoretical properties and is convenient for algorithmic developments.

Definition 2.2 (Goldstein δ-subdifferential). *Given a point x and $\delta\ge 0$, the Goldstein δ-subdifferential of Lipschitz f at x is defined by*

$$\partial_\delta f(x) := \mathrm{Co}\left\{\textstyle\bigcup_{y\in\mathbb{B}_\delta(x)}\partial f(y)\right\}.$$

We record some useful properties of the Clarke subdifferential and its Goldstein approximation here:

Fact 2.3 (cf. Clarke (1990); Goldstein (1977); Zhang et al. (2020)). *For an L-Lipschitz continuous f and $\delta>0$,*

- *$\partial f(x), \partial_\delta f(x)$ are nonempty, convex, compact;*
- *$\partial f(x) = \cap_{\delta>0}\cup_{y\in\mathbb{B}_\delta(x)}\partial f(y)$;*
- *$\partial f(x) = \cap_{\delta>0}\partial_\delta f(x)$;*
- *if f is C^1 near x, then $\partial f(x) = \{\nabla f(x)\}$;*
- *if f is convex, then $\partial f(x)$ is the convex subdifferential.*

We are now ready to introduce two important approximate stationarity notions. We refer the reader to (Davis & Drusvyatskiy, 2020) for a nice expository material.

Definition 2.4 (Goldstein approximate stationarity, GAS). *Given a locally Lipschitz function $f:\mathbb{R}^d\to\mathbb{R}$, we say that $x\in\mathbb{R}^d$ is an (ϵ,δ)-GAS point if*

$$\mathrm{dist}\Big(0, \partial_\delta f(x)\Big)\le\epsilon.$$

Definition 2.5 (near-approximate stationarity, NAS). *Given a locally Lipschitz function $f:\mathbb{R}^d\to\mathbb{R}$, we say that $x\in\mathbb{R}^d$ is an (ϵ,δ)-NAS point if*

$$\mathrm{dist}\Big(0, \textstyle\bigcup_{y\in\mathbb{B}_\delta(x)}\partial f(y)\Big)\le\epsilon.$$

It is easy to see that if x is NAS, then x is also GAS as $\partial_\delta f(x)\supseteq\cup_{y\in\mathbb{B}_\delta(x)}\partial f(y)$. But the converse does not hold in general, even for continuously differentiable functions.

Fact 2.6 (Kornowski & Shamir (2021, Proposition 1)). *For any $\delta>0$, there exists a continuously differentiable function $f:\mathbb{R}^2\to\mathbb{R}$, which is 2π-Lipschitz on $\delta\mathbb{B}$, such that $(0,0)$ is $(0,\delta)$-GAS but $\min_{x\in\delta\mathbb{B}}\|\nabla f(x)\|\ge 1$.*

Fact 2.6 does not hold for ρ-weakly convex functions with sufficiently small δ. Thus, it is still unclear whether NAS and GAS are equivalent assuming ρ-weak convexity with finite $\rho\ge 0$. We report below a convex polyhedral version (recall that convexity is 0-weak convexity), which might be of independent interest.

Proposition 2.7 (convex polyhedron). *For any $\delta>0$, there exists a convex function $f:\mathbb{R}^2\to\mathbb{R}$, which is 2-Lipschitz with polyhedral ∂f, such that $(0,2\delta)$ is $(0,\delta)$-GAS but $\min_{y\in\mathbb{B}_\delta((0,2\delta))}\mathrm{dist}\big(0,\partial f(y)\big)\ge\frac{2}{5}\sqrt{5}$.*

3. Computing GAS with Practical Oracle by Random Conic Perturbation

3.1. Subgradient Oracles

Assumption 3.1 (practical oracle). *Given x and Lipschitz continuous f:*

*(a) In the **deterministic** setting, if f is differentiable at x, then the oracle $\mathbb{O}_{\mathrm{d}}(x)$ returns a function value $f(x)$ and the gradient $\nabla f(x)$. Otherwise, it sets `error`$=1$.*

*(b) In the **stochastic** setting, if f is differentiable at x, then the oracle $\mathbb{O}_{\mathrm{s}}(x)$ returns a stochastic gradient g_x with $\mathbb{E}[g_x\mid\sigma(x)] = \nabla f(x)$ satisfying $\mathbb{E}[\|g_x-\nabla f(x)\|^2\mid\sigma(x)]\le\sigma^2$. Otherwise, it sets `error` $=1$.*

Compared with the oracle in (Zhang et al., 2020, Assumption 1), Assumption 3.1 only needs to evaluate the gradient $\nabla f(x)$ at differentiable points. Indeed, many modern Automatic Differentiation software (e.g., PyTorch, TensorFlow) can be used as an implementation of Assumption 3.1 without worrying about their incorrectness on subgradient evaluation for nonconvex nonsmooth function (Kakade & Lee, 2018).

3.2. Deterministic Setting

In this section, we present a practical algorithm for computing an (ϵ,δ)-GAS point and establish its finite-time complexity in the deterministic setting. The new algorithm replaces the stringent oracle assumption in (Zhang et al., 2020, Assumption 1(a)) with Assumption 3.1(a).

3.2.1. Algorithm

The main idea is to make use of the almost everywhere differentiability of Lipschitz functions as guaranteed by

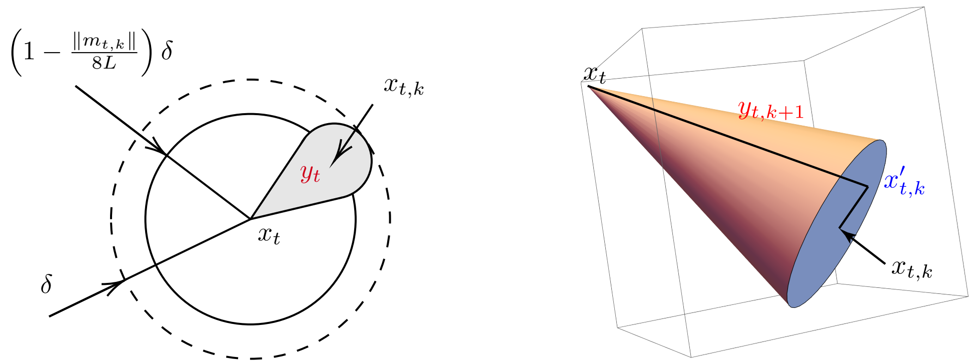

Figure 1. Random Conic Perturbation Scheme in Algorithm 1.

Rademacher's Theorem. By adopting a random conic perturbation to the uniform sampling direction in (Zhang et al., 2020, Algorithm 1), we have the following Algorithm 1, in which the main differences with (Zhang et al., 2020, Algorithm 1) are marked in blue. See also Figure 1.

Algorithm 1 Perturbed INGD

Input: $x_1 \in \mathbb{R}^d$.

Set $\texttt{error} = 0$.
for $t \in [T]$ **do**
  while $\|m_{t,K}\| > \epsilon$ **do**
    Sample $y_{t,1}$ uniformly from $\mathbb{B}_\delta(x_t)$.
    Call oracle $\sim, m_{t,1} = \mathbb{O}_{\mathrm{d}}(y_{t,1})$.
    for $k \in [K]$ **do**
      $x_{t,k} = x_t - \left(1 - \frac{\|m_{t,k}\|}{8L}\right) \cdot \delta \frac{m_{t,k}}{\|m_{t,k}\|}$.
      if $\|m_{t,k}\| \leq \epsilon$ **then**
        Terminate the algorithm and return x_t.
      else if $f(x_{t,k}) - f(x_t) < -\frac{\delta}{4}\|m_{t,k}\|$ **then**
        Set $x_{t+1} = x_{t,k}$ and $t = t + 1$.
        Break while-loop.
      else
        Sample $u_{t,k+1} \in \mathbb{R}^{d+1}$ uniformly from $\mathbb{S}^d$.
        Let $v_{t,k+1}$ be the first d coordinates of $u_{t,k+1}$.
        $b_{t,k+1} = v_{t,k+1} - \frac{v_{t,k+1}^\top (x_t - x_{t,k})}{\|x_t - x_{t,k}\|^2} \cdot (x_t - x_{t,k})$.
        Sample $y_{t,k+1}$ uniformly from $[x_t, x'_{t,k}]$, where $x'_{t,k} := x_{t,k} + \frac{\delta\|m_{t,k}\|}{8L} \cdot b_{t,k+1}$.
        Call oracle $\sim, g_{t,k+1} = \mathbb{O}_{\mathrm{d}}(y_{t,k+1})$.
        Update $m_{t,k+1} = \beta_{t,k} m_{t,k} + (1 - \beta_{t,k}) g_{t,k+1}$ with $\beta_{t,k} = \frac{8L^3 - L^2\|m_{t,k}\| - 4L\|m_{t,k}\|^2}{8L^3 - L^2\|m_{t,k}\| - \|m_{t,k}\|^3}$.
      end if
    end for
  end while
end for

3.2.2. Finite-Time Analysis

The main technical contributions in the analysis are summarized in the following two lemmas.

Lemma 3.2. *Let* $D := \{x : f \text{ is differentiable at } x\}$. *Given locally Lipschitz continuous* f, *we have*

$$\mathbb{P}\Big(\exists (t,k) \in [T] \times [K] : y_{t,k} \in D^c\Big) = 0.$$

Lemma 3.3. *Let* $K = \frac{80L^2}{\epsilon^2}$. *Given* $t \in [T]$, *it holds*

$$\mathbb{E}\Big[\|m_{t,K}\|^2\Big] \leq \frac{\epsilon^2}{16},$$

where $m_{t,k} = 0$ *for all* $k > k_0$ *if the* k*-loop breaks at* (t, k_0). *Consequently, for any* $0 \leq \gamma < 1$, *with probability* $1 - \gamma$, *there are at most* $\log(\gamma^{-1})$ *restarts of the while loop in the* t*-th iteration.*

We have the following finite-time guarantee for Algorithm 1.

Theorem 3.4. *Let* f *be* L*-Lipschitz continuous. Then, Algorithm 1 with* $K = \frac{80L^2}{\epsilon^2}$ *and* $T = \frac{4\Delta}{\epsilon\delta}$ *finds an* (ϵ, δ)*-GAS point with probability* $1 - \gamma$ *using at most*

$$\frac{320\Delta L^2}{\epsilon^3\delta} \log\left(\frac{4\Delta}{\gamma\epsilon\delta}\right) \qquad \text{oracle calls}$$

with $\mathbb{P}(\texttt{error} = 1) = 0$, *where* $f(x_0) - \inf_x f(x) \leq \Delta$.

3.3. Stochastic Setting

In this section, we consider the stochastic setting. The new algorithm replaces the stringent oracle assumption in (Zhang et al., 2020, Assumption 1(b)) with Assumption 3.1(b).

3.3.1. Algorithm

Technically speaking, the main difference from (Zhang et al., 2020, Algorithm 2) lies in the additional perturbation step. We need to carefully choose ζ to ensure that the iterates are within a δ-ball of some reference point without hurting the convergence. Since m_t is a weighted average of all the stochastic gradients, we need to show that it approximately belongs to the Goldstein δ-subdifferential $\partial_\delta f(x)$ of some reference point x.

The subtlety when $\|m_t\| = 0$: Unlike in the deterministic setting where we can terminate the algorithm if $\|m_{t,k}\|$ is small, in the stochastic case, m_t is a convex combination of stochastic gradients and thus it does not suffice to terminate the algorithm even if $\|m_t\| = 0$. The quantity that we aim to minimize is its expectation $\|\mathbb{E}[m_t]\| \leq \mathbb{E}[\|m_t\|]$. Due to this subtlety, we cannot let the perturbation size ζ_t adapt to $\|m_t\|$ as in the deterministic case: If $\zeta_t = \frac{\omega_1\|m_t\|}{p\|m_t\|+\omega_2}$ in Algorithm 2, then when $\|m_t\| = 0$, we have $y_{t+1} = x_{t+1} = x_t$, and we cannot ensure that f is differentiable at x_t almost surely. We choose a constant $\zeta_t \equiv \zeta$ in Algorithm 2 instead. In this case, when $\|m_t\| = 0$, y_{t+1} is sampled from a ball centered at x_t.

By adopting a random conic perturbation to (Zhang et al., 2020, Algorithm 2), we have the following Algorithm 2, in which the main differences with (Zhang et al., 2020, Algorithm 2) are marked in blue.

Algorithm 2 Perturbed Stochastic INGD

Input: $x_1 \in \mathbb{R}^d$.

Initialize: $m_1 = g_1 = \mathbb{O}_s(x_1)$. Set $\beta = 1 - \frac{\epsilon^2}{64G^2}$, $K = \frac{1}{\ln\frac{1}{\beta}} \ln\frac{16G}{\epsilon}$, $\omega = \left(\frac{1}{1-\beta} - \frac{1}{\ln\frac{1}{\beta}}\right)\ln\frac{16G}{\epsilon}$, $p = \frac{64G^2}{\delta\epsilon^2}\ln\frac{16G}{\epsilon}$, $q = \frac{256G^3}{\delta\epsilon^2}\ln\frac{16G}{\epsilon}$, $T = \frac{2^{16}G^3\Delta\ln\frac{16G}{\epsilon}}{\epsilon^4\delta}\max\{1, \frac{G\delta}{8\Delta}\}$.

Set $\texttt{error} = 0$.
for $t \in [T]$ **do**
  $x_{t+1} = x_t - \eta_t m_t$, where $\eta_t = \frac{1}{p\|m_t\|+q}$.
  Sample $u_{t+1} \in \mathbb{R}^{d+1}$ uniformly from $\mathbb{S}^d$.
  Let $v_{t+1} \in \mathbb{R}^d$ be the first d coordinates of u_{t+1}.
  If $\|m_t\| > 0, b_{t+1} = v_{t+1} - \frac{\langle v_{t+1}, x_t - x_{t+1}\rangle}{\|x_t - x_{t+1}\|^2}\cdot(x_t - x_{t+1})$; otherwise, $b_{t+1} = v_{t+1}$.
  Sample y_{t+1} uniformly from $[x_t, x_{t+1} + \zeta b_{t+1}]$, where $\zeta = \min\{\frac{\omega}{p}, \frac{\epsilon^2}{510q(L+G)}\}$.
  Call oracle $g_{t+1} = \mathbb{O}_s(y_{t+1})$.
  $m_{t+1} = \beta m_t + (1-\beta)g_{t+1}$.
end for

Output: $x_{\text{out}} := x_{\max\{1, i-K\}}$, where $i \sim \text{Unif}([T])$.

3.3.2. Finite-Time Analysis

We have the following finite-time guarantee for Algorithm 2, which is similar to (Zhang et al., 2020, Theorem 10) but replaces the stringent oracle assumption in (Zhang et al., 2020, Assumption 1(b)) with Assumption 3.1(b).

Theorem 3.5. *Under Assumption 3.1(b), with probability at least $\frac{3}{5}$, the output of Algorithm 2 satisfies* $\text{dist}(0, \partial_\delta f(x_{\text{out}})) \leq \epsilon$ *after at most*

$$\widetilde{O}\left(\frac{G^3\Delta}{\epsilon^4\delta}\right) \qquad \text{oracle calls}$$

with $\mathbb{P}(\texttt{error} = 1) = 0$, where $f(x_0) - \inf_x f(x) \leq \Delta$.

4. Computing NAS by GAS

In this section, we isolate a new function class within which the new algorithmic scheme can compute near-approximately stationary points in finite time. The new class goes far beyond that of ρ-weakly convex functions. We will first introduce the general results, and then several useful calculus rules. In Section 5, we will discuss applications of the new techniques to modern machine learning models.

4.1. General Results

The main strategy is to compute NAS by GAS. To this end, we need certain continuity of set-valued subdifferential mapping $\partial f : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$, which should be stronger than upper semicontinuity. A classic notion in set-valued analysis named outer Lipschitz continuity is defined as follows.

Definition 4.1 (Dontchev & Rockafellar (2009, 3D)). *A set-valued mapping $G : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is outer Lipschitz continuous (OLC) at $\bar{y}$ relative to a set D if $\bar{y} \in D \subset \text{dom}\, G$, $G(\bar{y})$ is a closed set, and there is a constant $\kappa \geq 0$ along with a neighborhood V of $\bar{y}$ such that*

$$G(y) \subseteq G(\bar{y}) + \kappa\|y - \bar{y}\|\mathbb{B}, \qquad \forall y \in V \cap D.$$

OLC is a weaker notion than Lipschitz continuity even for a single-valued mapping $G : \mathbb{R} \to \mathbb{R}$. See (Lewis & Pang, 2010, Example 2.4(a)). However, for our purposes, OLC is not sufficient since by the classic result of Robinson (1981) the bad function in Proposition 2.7 is OLC.

The following modified OLC notion for set-valued mapping is new and central in our development, which allows us to have a Lipschitz-type control of $G : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ from above within a constant-size neighborhood (see also Figure 2).

Definition 4.2 ((δ, η, κ)-outer Lipschitz continuous). *A set-valued mapping $G : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is (δ, η, κ)-OLC on S if for any $x \in S$, there exists a **pivot** $y \in \mathbb{B}_\delta(x) \cap S$ such that G is κ-OLC on $\mathbb{B}_\eta(x) \cap S$. In other words, for all $x \in S$, there exists a $y \in \mathbb{B}_\delta(x) \cap S$ such that*

$$G(z) \subseteq G(y) + \kappa\|y - z\|\mathbb{B}, \ \forall z \in \mathbb{B}_\eta(x) \cap S.$$

*Besides, we call $P^G : x \to y$ the **pivot mapping** of G.*

Remark 4.3. *A natural question about Definition 4.2 is why we want to set η and δ to different values. In other words, why (δ, δ, κ)-OLC is not sufficient. Consider the convex function $g(x, y) := \max\{2x, -2x, y\}$, which is the bad function in the proof of Proposition 2.7. It is easy to see that ∂g is polyhedral and OLC by (Robinson, 1981). However, for any $\delta > 0$, ∂g is not (δ, δ, κ)-OLC at $(0, 2\delta), \forall \kappa \geq 0$. Thus, even for an OLC mapping ∂g, we cannot promise $\exists \delta > 0$ such that ∂g is (δ, δ, κ)-OLC at certain $x \in \text{dom}\, \partial g$. Instead, we will show in Theorem 4.5 that if ∂g is OLC*

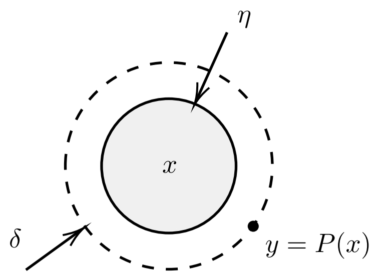

Figure 2. (δ, η, κ)-Outer Lipschitz Continuity in Definition 4.2.

and S is compact, then $\forall \delta > 0, \exists \eta > 0$ such that ∂g is (δ, η, κ)-OLC on S.

We are now ready for the main theorem of this section:

Theorem 4.4 (NAS by GAS)**.** *For a Lipschitz continuous f, suppose that $\partial f : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is (δ, η, κ)-OLC. If x is (ϵ, η)-GAS, then x is $\big(\epsilon + \kappa(\delta + \eta), \delta\big)$-NAS.*

It is natural to ask what function class admits a (δ, η, κ)-OLC subdifferential.

Theorem 4.5. *Let $\delta > 0$ and $\partial f : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be κ-OLC. For any compact set S, there exists an $\eta \in (0, \delta]$ such that ∂f is (δ, η, κ)-OLC on S.*

Remark 4.6. *If the set of (ϵ, δ)-GAS points is bounded and ∂f is OLC, then we can use Theorem 4.5 and Theorem 4.4 to guarantee NAS from GAS. Note that functions with OLC subdifferential have been widely studied in the variational analysis literature. For example, ∂f with a finite union of convex polyhedral graph (Robinson, 1981) is OLC. If f is an ℓc-stable function (Bednařík & Pastor, 2013, Theorem 2), then ∂f is OLC.*

Given an OLC mapping ∂f and a constant $\delta > 0$, it is generally hard to estimate the constant η as guaranteed by Theorem 4.5. However, its value is needed for the stopping rules of Algorithm 1. In the next subsection, we provide several useful calculus rules to compute the parameter η explicitly.

4.2. Calculus of (δ, η, κ)-Outer Lipschitz Continuity

In this section, we establish a series of calculus rules to verify and compute the parameters in Definition 4.2. We first introduce four rules that have taken the subdifferential calculus rules[4] of f into consideration.

Proposition 4.7 (smooth regularization)**.** *Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ has a (δ, η, κ)-OLC ∂f and $g : \mathbb{R}^d \to \mathbb{R}$ is differentiable with a β-Lipschitz gradient ∇g. Then $\partial(f + g)$ is $(\delta, \eta, \beta + \kappa)$-OLC.*

[4] We note here that the validity of subdifferential chain rules for nonconvex nonsmooth functions is highly non-trivial. See, for example, (Rockafellar & Wets, 2009, Chapter 10).

Proposition 4.8 (separable sum)**.** *Suppose, for any $i \in [m]$, that $f_i : \mathbb{R}^{d_i} \to \mathbb{R}$ has a $(\delta_i, \eta_i, \kappa_i)$-OLC ∂f_i. Let $f(x) := \sum_{i=1}^m f_i(x_i)$, where $x := \bigoplus_{i=1}^m x_i$. Then, ∂f is (δ, η, κ)-OLC with*

$$\delta = \sqrt{\sum_{i=1}^m \delta_i^2}, \quad \eta = \min_{i \in [m]} \eta_i, \quad \kappa = \sqrt{\sum_{i=1}^m \kappa_i^2}.$$

Proposition 4.9 (linear composition)**.** *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ has a (δ, η, κ)-OLC ∂f and $A \in \mathbb{R}^{n \times d}$ is surjective. Then, $\partial(f \circ A)$ is $\Big(\delta \|A^\dagger\|, \frac{\eta}{\|A\|}, \kappa \|A\|^2\Big)$-outer Lipschitz continuous.*

Proposition 4.10 (rescaling)**.** *Suppose that the L_1-Lipschitz $f : \mathbb{R}^n \to \mathbb{R}$ has a (δ, η, κ)-OLC ∂f and $g : \mathbb{R} \to \mathbb{R}$ is L_2-Lipschitz and β-smooth. Then, $\partial(g \circ f)$ is $(\delta, \eta, \beta L_1 + \kappa L_2)$-OLC.*

Then, we introduce a partial sum rule, which is powerful but needs to be used in conjunction with certain subdifferential calculus rules (e.g., assuming Clarke regularity). The following rules are crucial in the 2-layer ReLU neural network example (see Section 5.2.2).

Proposition 4.11 (sum)**.** *Suppose, for any $i \in [m]$, that $G_i : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is $(\delta_i, \eta_i, \kappa_i)$-OLC with a shared pivot mapping $P : \mathbb{R}^d \to \mathbb{R}^d$. Let $G(x) := \sum_{i=1}^m G_i(x)$. Then, G is (δ, η, κ)-OLC with*

$$\delta = \min_{i \in [m]} \delta_i, \quad \eta = \min_{i \in [m]} \eta_i, \quad \kappa = \sum_{i=1}^m \kappa_i.$$

Corollary 4.12 (partially separable sum)**.** *Suppose, for any $i \in [m]$, that $G_i : \mathbb{R}^{d_0} \times \mathbb{R}^{d_i} \rightrightarrows \mathbb{R}^{d_0} \times \mathbb{R}^{d_i}$ is $(\delta_i, \eta_i, \kappa_i)$-OLC with a partially shared pivot mapping $P_i : \mathbb{R}^{d_0} \times \mathbb{R}^{d_i} \to \mathbb{R}^{d_0} \times \mathbb{R}^{d_i}$, such that $\pi_1 \circ P_i(x_0, x_i) = \pi_1 \circ P_1(x_0, x_1), \forall i \in [m]$. Let $G(x) := \sum_{i=1}^m G_i(x_0, x_i)$, where $x := \bigoplus_{i=0}^m x_i$. Then, G is (δ, η, κ)-OLC with*

$$\delta = \sqrt{\sum_{i=1}^m \delta_i^2}, \quad \eta = \min_{i \in [m]} \eta_i, \quad \kappa = \sum_{i=1}^m \kappa_i.$$

4.3. Discussion

We record here a recipe to prove (δ, η, κ)-OLC from scratch, which when combined with the calculus rules in this section forms a toolbox for determining the parameters (δ, η, κ).

S1. Construct pivot mapping $P : \mathbb{R}^d \to \mathbb{R}^d$.

S2. Verify $\|x - P(x)\| \le \delta$ for all $x \in \mathbb{R}^d$.

S3. Prove that for all $x \in \mathbb{R}^d$, it holds

$$G(z) \subseteq G(P(x)) + \kappa \|z - P(x)\| \mathbb{B}, \ \forall z \in \mathbb{B}_\eta(x) \cap S.$$

We will provide concrete examples in Section 5.

5. Applications

To demonstrate the wide applicability of the new theoretical framework, we discuss examples in machine learning, namely ρ-margin SVM, 1-layer, and 2-layer ReLU NN, all being Clarke irregular and not weakly convex. We show that all these examples are subdifferential (δ, η, κ)-OLC, where the parameters (δ, η, κ) can be determined via the calculus rules in Section 4.2.

5.1. ρ-Margin loss SVM

We aim to solve

$$\min_{w \in \mathbb{R}^d} F(w) := \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \phi_\rho(z_i^\top w), \qquad (\rho\text{-MSVM})$$

where $\phi_\rho(u) := \min\left(1, \max\left(0, 1 - \frac{u}{\rho}\right)\right)$.

The goal is to compute (ϵ, δ)-NAS points for Problem (ρ-MSVM) by computing (ϵ', δ')-GAS points. We note that the ρ-Margin loss SVM in Problem (ρ-MSVM) and its $\rho = 1$ version, also known as ramp loss SVM, have been widely recognized in the operations research (Brooks, 2011; Carrizosa et al., 2014; Wang et al., 2021; Tian & So, 2022), statistics (Shen et al., 2003; Wu & Liu, 2007; Liu et al., 2005), and machine learning (Huang et al., 2014; Keshet & McAllester, 2011; Collobert et al., 2006b;a; Ertekin et al., 2010; Suzumura et al., 2017; Maibing & Igel, 2015) communities as providing better robustness against data outliers than the vanilla SVM. The general ρ-version can be found in the learning theory textbook (Mohri et al., 2018, Corollary 5.11).

It is elementary to see that $\partial \phi_\rho$ is $(\delta, \delta, 0)$-OLC for any $0 < \delta \le \frac{\rho}{2}$ with pivot mapping $P^{\phi_\rho} : \mathbb{R} \to \mathbb{R}$ defined by

$$P^{\phi_\rho}(x) := \begin{cases} 0 & \text{for} \quad |x| \le \frac{\rho}{2}, \\ \rho & \text{for} \quad |x - \rho| < \frac{\rho}{2}, \\ x & \text{for} \quad \text{otherwise}. \end{cases}$$

Let $\Phi_\rho(y) := \sum_{i=1}^n \phi_\rho(y_i)$. Then, by Proposition 4.8, $\partial\Phi_\rho$ is $(\sqrt{n}\delta, \delta, 0)$-OLC. Assuming that $Z \in \mathbb{R}^{n\times d}$ is surjective, by Proposition 4.9, $\partial(\Phi_\rho \circ Z)$ is $\left(\sqrt{n}\delta\|Z^\dagger\|, \frac{\delta}{\|Z\|}, 0\right)$-OLC. Using Proposition 4.7, ∂F is $\left(\sqrt{n}\delta\|Z^\dagger\|, \frac{\delta}{\|Z\|}, 1\right)$-OLC. By Theorem 4.4, if x is $\left(\epsilon, \frac{\delta}{\|Z\|}\right)$-GAS, then it is also $\left(\epsilon + \left(\sqrt{n}\|Z^\dagger\| + \frac{1}{\|Z\|}\right)\delta, \sqrt{n}\|Z^\dagger\|\delta\right)$-NAS. Let the condition number of Z be $\kappa(K) := \|Z^\dagger\|\|Z\|$. In other words, to compute an (ϵ, δ)-NAS point, it is sufficient to have an (ϵ', δ')-GAS point, where (in a dimension-free manner)

$$\epsilon' \le \frac{\epsilon}{2} \text{ and } \delta' \le \min\left\{\frac{\delta}{\sqrt{n}\kappa(Z)}, \frac{\epsilon}{2\sqrt{n}\kappa(Z) + 2}, \frac{\rho}{2\|Z\|}\right\}.$$

5.2. Shallow ReLU Neural Network.

In this subsection, we will discuss the computation of (ϵ, δ)-NAS points for shallow ReLU neural networks. For simplicity, we will not trace explicitly the constants (δ, η, κ) in this subsection. Instead, we will say that f is *subdifferentially OLC trackable* if the parameters (δ, η, κ) of ∂f can be determined by the calculus rules in Section 4.2.

Recently, finite-time convergence of neural networks in the overparameterized regime has been extensively studied (Jacot et al., 2018; Chizat et al., 2019; Du et al., 2018; Arora et al., 2019; Du et al., 2019; Zou et al., 2020). For the underparameterized regime, the asymptotic convergence of ReLU neural network is analyzed in the continuous-time gradient flow sense (Eberle et al., 2021; Jentzen & Riekert, 2021). However, it is still unclear what convergence guarantee we can have for (potentially underparameterized) ReLU neural networks within finite time as they are not weakly convex and the finite-time analyses in (Davis & Grimmer, 2019; Davis & Drusvyatskiy, 2019) are inapplicable.

5.2.1. 1-Layer ReLU Neural Network

We first investigate the easy case, that is, the 2-layer ReLU neural network with the weights of the second layer fixed. It is notable that we will not impose any assumption on the number of hidden nodes m.

Let $\sigma(u) := \max\{u, 0\}$. Setting pivot $y := \mathbb{1}_{|u|>\delta} u$, it is elementary to see that $\partial\sigma : \mathbb{R} \rightrightarrows \mathbb{R}$ is $(\delta, \delta, 0)$-OLC for any $\delta > 0$. Similarly, $\partial(-\sigma)$ is $(\delta, \delta, 0)$-OLC for any $\delta > 0$. We aim to solve

$$\min_{W \in \mathbb{R}^{d\times m}} F(W) := \sum_{i=1}^{n} \ell\left(y_i, \sum_{j=1}^{m} (-1)^j \sigma\left(w_j^\top x_i\right)\right) + R(W).$$

Suppose that the regularization term $R : \mathbb{R}^{d\times m} \to \mathbb{R}$ is smooth. Let $h : \mathbb{R}^m \to \mathbb{R}$ be given by $h(u) := \sum_{j=1}^m (-1)^j \sigma(u_j)$. By Proposition 4.8, h is subdifferentially OLC trackable. Let $\ell_i(u) = \ell(y_i, u)$ and assume that ℓ_i is Lipschitz and smooth. Let $f : \mathbb{R}^{mn} \to \mathbb{R}$ be given by $f(\mathrm{Vec}(U)) := \sum_{i=1}^n \ell_i \circ h(u_i)$, where $u_i \in \mathbb{R}^m, \forall i \in [n]$ and $U \in \mathbb{R}^{m\times n}$. With Proposition 4.10 and using Proposition 4.8 again, f is subdifferentially OLC trackable. We assume that the data $X \in \mathbb{R}^{n\times d}$ is surjective, which holds in many modern high-dimensional machine learning scenarios. Let $x_i \in \mathbb{R}^{1\times d}$ be the i-th row of X. We define $X_i \in \mathbb{R}^{m\times md}$ and $X_{\text{big}} \in \mathbb{R}^{mn\times md}$ as

$$X_i := \begin{bmatrix} x_i & & & \\ & x_i & & \\ & & \ddots & \\ & & & x_i \end{bmatrix}, \qquad X_{\text{big}} := \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}.$$

As X is surjective, X_{big} is surjective. Using Proposition 4.9, we have $f(X_{\text{big}}\,\mathrm{Vec}(W))$ is subdifferentially

OLC trackable, where $\mathrm{Vec}(W) \in \mathbb{R}^{md}$. By $F(W) = f(X_{\text{big}}\,\mathrm{Vec}(W)) + R(W)$ and Proposition 4.7, $F(W)$ is subdifferentially OLC trackable.

5.2.2. 2-Layer ReLU Neural Network

Let $\varrho(a,b) := a \cdot \max\{b, 0\}$. We aim to solve

$$\min_{\substack{W\in\mathbb{R}^{d\times m}\\ a\in\mathbb{R}^m}} F(W,a) := \sum_{i=1}^n \ell\left(y_i, \sum_{j=1}^m \varrho(a_j, w_j^\top x_i)\right) + R(W,a)$$

with surjective $X \in \mathbb{R}^{n\times d}$ and smooth regularization term $R : \mathbb{R}^{d\times m} \times \mathbb{R}^m \to \mathbb{R}$.

Compared with the 1-layer case, the main difficulty in the analysis is due to the inseparability of $\{a_j\}_{j\in[m]}$, as one cannot apply the subdifferential chain rule and OLC calculus rules directly. To cope with this, we need the partial separable rule in Corollary 4.12 and a partially differentiable sum rule in Proposition 5.2, which might be of independent interest. To begin, we have the following subdifferential characterization of $\partial\varrho : \mathbb{R}^2 \rightrightarrows \mathbb{R}^2$:

Claim 5.1. *For $\varrho(u_1,u_2) := u_1 \cdot \max\{u_2, 0\}$, it holds*

$$\partial\varrho(u_1,u_2) = \begin{cases} (u_2,u_1) & \textit{for} \quad u_2 > 0, \\ (0,0) & \textit{for} \quad u_2 < 0, \\ (0, \mathrm{Co}\{0, u_1\}) & \textit{for} \quad u_2 = 0. \end{cases}$$

Then, we investigate the continuity of $\partial\varrho$. Given any $\delta > 0$, $x \in \mathbb{R}^2$, and $z \in \mathbb{B}_\delta(x)$, we consider the following cases:

- If $|x_2| > \delta$, let $y = x$.
  - If $y_2 > 0$, then $z_2 > 0$. We have $\partial\varrho(z) = (z_2, z_1) \subseteq (y_2, y_1) + \|y-z\|\mathbb{B} = \partial\varrho(y) + \|y-z\|\mathbb{B}$.
  - If $y_2 < 0$, then $z_2 < 0$. We have $\partial\varrho(z) = (0,0) = \partial\varrho(y)$.
- If $0 \le |x_2| \le \delta$, let $y = (x_1, 0)$. It is easy to see that $\|y - x\| = |x_2| \le \delta$.
  - If $z_2 > 0$, we have $\partial\varrho(z) = (z_2, z_1) \subseteq (0, y_1) + \|y-z\|\mathbb{B} \subseteq \partial\sigma(y) + \|y-z\|\mathbb{B}$.
  - If $z_2 < 0$, we have $\partial\varrho(z) = (0,0) \subseteq (0,0) + \|y-z\|\mathbb{B} \subseteq \partial\sigma(y) + \|y-z\|\mathbb{B}$.
  - If $z_2 = 0$, we have $\partial\varrho(z) = (0, \mathrm{Co}\{0, z_1\}) \subseteq (0, \mathrm{Co}\{0, y_1\}]) + \|y-z\|\mathbb{B} = \partial\sigma(y) + \|y-z\|\mathbb{B}$.

Therefore, for any $\delta > 0$, $\partial\varrho$ is $(\delta,\delta,1)$-OLC with pivot mapping $P^{\partial\varrho} : \mathbb{R}^2 \to \mathbb{R}^2$ defined by

$$P^{\partial\varrho}\big((x_1,x_2)\big) := \begin{cases} (x_1, 0) & \text{for} \quad |x| \le \delta, \\ (x_1, x_2) & \quad\quad \text{otherwise}. \end{cases}$$

It is easy to see that $\pi_1 \circ P^{\partial\varrho}\big((x_1,x_2)\big)$ is independent of x_2. Let $h_i : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}, \forall i \in [n]$ be defined by

$$h_i(a, u_i) := \ell_i\left(\sum_{j=1}^m \varrho(a_j, u_{ij})\right).$$

Then, by the choices of pivots in the proof of Propositions 4.10 and 4.11, ∂h_i is subdifferentially OLC trackable with pivot mapping $P^{\partial h_i} : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^m \times \mathbb{R}^m$ defined by $P^{\partial h_i}\big((a,u_i)\big) := (a, \tilde{u}_i)$, where

$$\tilde{u}_{ij} := \begin{cases} 0 & \text{for} \quad |u_{ij}| \le \delta, \\ u_{ij} & \quad\quad \text{otherwise}. \end{cases}$$

Therefore, $\{\partial h_i\}_{i\in[m]}$ partially shares the pivot mapping $P^{\partial h_i}$ on the first argument, i.e., $\pi_1 \circ P^{\partial h_i}\big((x_0, x_i)\big) = \pi_1 \circ P^{\partial h_1}\big((x_0,x_1)\big), \forall i \in [n]$. Let $f(a,U) := \sum_{i=1}^n h_i(a, u_i)$. By Corollary 4.12, $\sum_{i=1}^n \partial h_i$ is subdifferentially OLC trackable. To proceed, we need the following chain rule, whose proof is technical and might be of independent interest.

Proposition 5.2 (partially differentiable sum rule)**.** *It holds*

$$\partial f(a,U) = \sum_{i=1}^n \partial h_i(a, u_i).$$

Then, ∂f is subdifferentially OLC trackable. Suppose that the data $X \in \mathbb{R}^{n\times d}$ is surjective. Let x_i be the i-th row of X. We define $\theta \in \mathbb{R}^{m+md}$ and $X_{\text{huge}} \in \mathbb{R}^{(m+mn)\times(m+md)}$ as

$$\theta := \begin{bmatrix} a \\ \mathrm{Vec}(W) \end{bmatrix}, \qquad X_{\text{huge}} := \begin{bmatrix} I_m & \\ & X_{\text{big}} \end{bmatrix}.$$

As X is surjective, X_{huge} is surjective. Using Proposition 4.9, we have $f(X_{\text{huge}}\theta)$ is subdifferentially OLC trackable. By $F(W,a) = f(X_{\text{huge}}\theta) + R(W,a)$ and Proposition 4.7, $F(W,a)$ is subdifferentially OLC trackable.

6. Closing Remarks

In this paper, we report a practical algorithmic scheme to compute GAS points for general Lipschitz functions with finite-time complexity. We also isolate a new function class for which our scheme computes NAS points in finite time. Besides, we establish a series of theoretical tools to compute parameters in our new function class. To demonstrate the wide applicability of our new theoretical framework, we discuss modern machine learning models and show that they satisfy our new conditions. We hope that our results can be beneficial to the understanding of finite-time complexity of sharper approximate stationarity for Lipschitz continuous "non"-problems. An intriguing further direction is to apply the new analytical framework to other nonconvex nonsmooth problems. Extending the calculus rules in Section 4.2 or refining the modified OLC notion in Definition 4.2 would also be interesting.

References

Arora, S., Du, S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pp. 322–332, 2019.

Barthe, F., Guédon, O., Mendelson, S., and Naor, A. A probabilistic approach to the geometry of the ℓ_p^n-ball. *The Annals of Probability*, 33(2):480–513, 2005.

Beck, A. and Hallak, N. On the convergence to stationary points of deterministic and randomized feasible descent directions methods. *SIAM Journal on Optimization*, 30 (1):56–79, 2020.

Bednařík, D. and Pastor, K. On lipschitz behaviour of some generalized derivatives. *Mathematica Slovaca*, 63(3): 587–594, 2013.

Benaïm, M., Hofbauer, J., and Sorin, S. Stochastic approximations and differential inclusions. *SIAM Journal on Control and Optimization*, 44(1):328–348, 2005.

Bolte, J., Sabach, S., Teboulle, M., and Vaisbourd, Y. First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM Journal on Optimization*, 28(3):2131–2151, 2018.

Brooks, J. P. Support vector machines with the ramp loss and the hard margin loss. *Operations Research*, 59(2): 467–479, 2011.

Burke, J. V., Lewis, A. S., and Overton, M. L. Approximating subdifferentials by random sampling of gradients. *Mathematics of Operations Research*, 27(3):567–584, 2002.

Burke, J. V., Lewis, A. S., and Overton, M. L. A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. *SIAM Journal on Optimization*, 15(3):751–779, 2005.

Burke, J. V., Curtis, F. E., Lewis, A. S., Overton, M. L., and Simões, L. E. Gradient sampling methods for nonsmooth optimization. *Numerical Nonsmooth Optimization: State of the Art Algorithms*, pp. 201–225, 2020.

Carrizosa, E., Nogales-Gómez, A., and Morales, D. R. Heuristic approaches for support vector machines with the ramp loss. *Optimization Letters*, 8(3):1125–1135, 2014.

Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32:2937–2947, 2019.

Clarke, F. H. *Optimization and Nonsmooth Analysis*. SIAM, 1990.

Collobert, R., Sinz, F., Weston, J., and Bottou, L. Trading convexity for scalability. In *International Conference on Machine Learning*, pp. 201–208, 2006a.

Collobert, R., Sinz, F., Weston, J., Bottou, L., and Joachims, T. Large scale transductive SVMs. *Journal of Machine Learning Research*, 7(8), 2006b.

Daniilidis, A. and Drusvyatskiy, D. Pathological subgradient dynamics. *SIAM Journal on Optimization*, 30(2): 1327–1338, 2020.

Davis, D. and Drusvyatskiy, D. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

Davis, D. and Drusvyatskiy, D. Subgradient methods under weak convexity and tame geometry. *SIAG/OPT Views and News*, 28(1), 2020.

Davis, D. and Grimmer, B. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *SIAM Journal on Optimization*, 29(3):1908–1930, 2019.

Davis, D., Drusvyatskiy, D., Kakade, S., and Lee, J. D. Stochastic subgradient method converges on tame functions. *Foundations of Computational Mathematics*, 20(1): 119–154, 2020.

Davis, D., Drusvyatskiy, D., Lee, Y. T., Padmanabhan, S., and Ye, G. A gradient sampling method with complexity guarantees for lipschitz functions in high and low dimensions. *arXiv preprint arXiv:2112.06969*, 2021.

Dontchev, A. L. and Rockafellar, R. T. *Implicit Functions and Solution Mappings*, volume 543. Springer, 2009.

Drusvyatskiy, D. and Paquette, C. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178(1):503–558, 2019.

Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pp. 1675–1685, 2019.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2018.

Duchi, J. C. and Ruan, F. Stochastic methods for composite and weakly convex optimization problems. *SIAM Journal on Optimization*, 28(4):3229–3259, 2018.

Eberle, S., Jentzen, A., Riekert, A., and Weiss, G. S. Existence, uniqueness, and convergence rates for gradient flows in the training of artificial neural networks with ReLU activation. *arXiv preprint arXiv:2108.08106*, 2021.

Ertekin, S., Bottou, L., and Giles, C. L. Nonconvex online support vector machines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2):368–381, 2010.

Goldstein, A. Optimization of Lipschitz continuous functions. *Mathematical Programming*, 13(1):14–22, 1977.

Huang, X., Shi, L., and Suykens, J. A. Ramp loss linear programming support vector machine. *Journal of Machine Learning Research*, 15(1):2185–2211, 2014.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, pp. 8580–8589, 2018.

Jentzen, A. and Riekert, A. On the existence of global minima and convergence analyses for gradient descent methods in the training of deep neural networks. *arXiv preprint arXiv:2112.09684*, 2021.

Kakade, S. M. and Lee, J. D. Provably correct automatic subdifferentiation for qualified programs. In *Advances in Neural Information Processing Systems*, pp. 7125–7135, 2018.

Keshet, J. and McAllester, D. Generalization bounds and consistency for latent structural probit and ramp loss. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

Khan, K. A. and Barton, P. I. Evaluating an element of the clarke generalized jacobian of a composite piecewise differentiable function. *ACM Transactions on Mathematical Software*, 39(4):1–28, 2013.

Kiwiel, K. C. Convergence of the gradient sampling algorithm for nonsmooth nonconvex optimization. *SIAM Journal on Optimization*, 18(2):379–388, 2007.

Kiwiel, K. C. A nonderivative version of the gradient sampling algorithm for nonsmooth nonconvex optimization. *SIAM Journal on Optimization*, 20(4):1983–1994, 2010.

Kornowski, G. and Shamir, O. Oracle complexity in nonsmooth nonconvex optimization. In *Advances in Neural Information Processing Systems*, volume 34, pp. 324–334, 2021.

Lewis, A. S. and Pang, C. J. Lipschitz behavior of the robust regularization. *SIAM Journal on Control and Optimization*, 48(5):3080–3104, 2010.

Liu, Y., Shen, X., and Doss, H. Multicategory ψ-learning and support vector machine: computational tools. *Journal of Computational and Graphical Statistics*, 14(1):219–236, 2005.

Maibing, S. F. and Igel, C. Computational complexity of linear large margin classification with ramp loss. In *International Conference on Artificial Intelligence and Statistics*, pp. 259–267. PMLR, 2015.

Majewski, S., Miasojedow, B., and Moulines, E. Analysis of nonsmooth stochastic approximation: The differential inclusion approach. *arXiv preprint arXiv:1805.01916*, 2018.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. MIT press, 2018.

Munkres, J. R. *Topology: A First Course*. Prentice-Hall, 1974.

Nemirovskij, A. S. and Yudin, D. B. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience, 1983.

Nesterov, Yu. Lexicographic differentiation of nonsmooth functions. *Mathematical Programming*, 104(2):669–700, 2005.

Robinson, S. M. Some continuity properties of polyhedral multifunctions. In *Mathematical Programming at Oberwolfach*, pp. 206–214. Springer, 1981.

Rockafellar, R. T. and Wets, R. J.-B. *Variational Analysis*, volume 317. Springer Science & Business Media, 2009.

Shen, X., Tseng, G. C., Zhang, X., and Wong, W. H. On ψ-learning. *Journal of the American Statistical Association*, 98(463):724–734, 2003.

Suzumura, S., Ogawa, K., Sugiyama, M., Karasuyama, M., and Takeuchi, I. Homotopy continuation approaches for robust SV classification and regression. *Machine Learning*, 106(7):1009–1038, 2017.

Tian, L. and So, A. M.-C. On the hardness of computing near-approximate stationary points of Clarke regular nonsmooth nonconvex problems and certain DC programs. *ICML Workshop on Beyond First-Order Methods in ML Systems*, 2021.

Tian, L. and So, A. M.-C. Computing d-stationary points of ρ-margin loss svm. In *International Conference on Artificial Intelligence and Statistics*, pp. 3772–3793. PMLR, 2022.

Vial, J.-P. Strong and weak convexity of sets and functions. *Mathematics of Operations Research*, 8(2):231–259, 1983.

Wang, H., Shao, Y., and Xiu, N. Proximal operator and optimality conditions for ramp loss SVM. *Optimization Letters*, pp. 1–16, 2021.

Wolfe, P. A method of conjugate subgradients for minimizing nondifferentiable functions. In *Nondifferentiable Optimization*, pp. 145–173. Springer, 1975.

Wu, Y. and Liu, Y. Robust truncated hinge loss support vector machines. *Journal of the American Statistical Association*, 102(479):974–983, 2007.

Zhang, J., Lin, H., Jegelka, S., Jadbabaie, A., and Sra, S. Complexity of finding stationary points of nonsmooth nonconvex functions. In *International Conference on Machine Learning*, pp. 11173–11182, 2020.

Zou, D., Cao, Y., Zhou, D., and Gu, Q. Gradient descent optimizes over-parameterized deep ReLU networks. *Machine Learning*, 109(3):467–492, 2020.

A. Proofs of Section 2

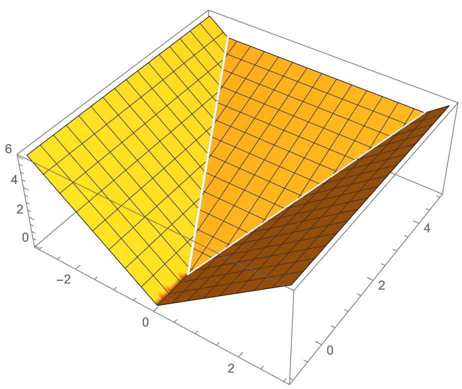

Figure 3. The function used in the proof of Proposition 2.7.

Proposition 2.7 (convex polyhedron)**.** *For any $\delta > 0$, there exists a convex function $f : \mathbb{R}^2 \to \mathbb{R}$, which is 2-Lipschitz with polyhedral ∂f, such that $(0, 2\delta)$ is $(0, \delta)$-GAS but $\min_{y \in \mathbb{B}_\delta((0,2\delta))} \operatorname{dist}\big(0, \partial f(y)\big) \geq \frac{2}{5}\sqrt{5}$.*

Proof. Fixing some $\delta > 0$, consider the function (see also Figure 3), whose convexity is obvious,

$$f(x, y) \coloneqq \max\{2x, -2x, y\}.$$

Note that, by $(\pm\delta, 2\delta) \in \mathbb{B}_\delta\big((0, 2\delta)\big)$, it holds

$$(0,0) = \frac{1}{2}(-2,0) + \frac{1}{2}(2,0) \in \frac{1}{2}\partial f\big((-\delta, 2\delta)\big) + \frac{1}{2}\partial f\big((\delta, 2\delta)\big) \subseteq \partial_\delta f\big((0, 2\delta)\big).$$

Besides, as $(0,0) \notin \mathbb{B}_\delta\big((0, 2\delta)\big)$, it is elementary to see

$$\operatorname{dist}\Big(0, \textstyle\bigcup_{y \in \mathbb{B}_\delta((0,2\delta))} \partial f(y)\Big) \geq \min_{0 \leq \lambda \leq 1} \|(2\lambda, 0) + (0, 1-\lambda)\| = \frac{2}{5}\sqrt{5},$$

as required. $\square$

B. Proofs of Section 3.2

Lemma 3.2. *Let $D \coloneqq \{x : f \text{ is differentiable at } x\}$. Given locally Lipschitz continuous f, we have*

$$\mathbb{P}\Big(\exists (t,k) \in [T] \times [K] : y_{t,k} \in D^c\Big) = 0.$$

Proof. Fix $(t,k) \in [T] \times [K]$. Let

$$S_1 \coloneqq \left\{(\lambda, \xi) : \lambda \in [0,1], \xi \in \mathbb{R}^{d-1}, \|\xi\| \leq 1\right\},$$
$$S_2 \coloneqq \left\{y \in \mathbb{R}^d : y = x_t + \lambda\left(x_{t,k} - x_t + \frac{\delta\|m_{t,k}\|}{8L} \cdot b_{t,k+1}\right), \lambda \in [0,1], \|b_{t,k+1}\| \leq 1, b_{t,k+1}^\top (x_{t,k} - x_t) = 0\right\}.$$

Let $X^\perp \in \mathbb{R}^{d \times d-1}$ be an orthonormal basis of $\operatorname{span}(x_{t,k} - x_t)^\perp$. We define the following isomorphism:

$$T : S_1 \longrightarrow S_2$$
$$(\lambda, \xi) \longrightarrow y_{t,k+1} \coloneqq x_t + \lambda\left(x_{t,k} - x_t + \frac{\delta\|m_{t,k}\|}{8L} \cdot X^\perp \xi\right).$$

Then, by Rademacher theorem (Rockafellar & Wets, 2009, Theorem 9.60) and T^{-1} is Lipschitz, we have

$$m\left(y \in D^c \cap S_2\right) = m\big((\lambda, \xi) \in T^{-1}(D^c \cap S_2)\big) = 0.$$

Let $S_3 \coloneqq \{b \in \mathbb{R}^d : \|b\| \le 1, b^\top (x_{t,k} - x_t) = 0\}$. By (Barthe et al., 2005, Corollary 4), we have $b_{t,k+1} \sim \mathrm{Unif}(S_3) \stackrel{d}{=} X^\perp \xi$, where $\xi \sim \mathrm{Unif}(\mathbb{B}^{d-1})$. With $\lambda \sim \mathrm{Unif}([0,1])$ and countable union of zero measure set is negligible, we have

$$\mathbb{P}\big(y_{t,k} \in D^c \cap S_2, \forall (t,k) \in [T] \times [K]\big) = 0,$$

which completes the proof. $\square$

Lemma 3.3. *Let $K = \frac{80L^2}{\epsilon^2}$. Given $t \in [T]$, it holds*

$$\mathbb{E}\Big[\|m_{t,K}\|^2\Big] \le \frac{\epsilon^2}{16},$$

where $m_{t,k} = 0$ for all $k > k_0$ if the k-loop breaks at (t, k_0). Consequently, for any $0 \le \gamma < 1$, with probability $1-\gamma$, there are at most $\log(\gamma^{-1})$ restarts of the while loop in the t-th iteration.

Proof. Let $\mathcal{F}_{t,k} = \sigma(y_{t,1}, \cdots, y_{t,k})$ and $\widehat{\mathcal{F}}_{t,k} = \sigma(y_{t,1}, \cdots, y_{t,k}, b_{t,k+1})$. We denote $D_{t,k}$ as the event that k-loop does not break at $x_{t,k}$, i.e., $\|m_{t,k}\| > \epsilon$ and $f(x_{t,k}) - f(x_t) > -\frac{\delta}{4}\|m_{t,k}\|$. It is clear that $D_{t,k} \in \mathcal{F}_{t,k} \subset \widehat{\mathcal{F}}_{t,k}$. Let $\gamma(\lambda) = (1-\lambda)x_t + \lambda\left(x_{t,k} + \frac{\delta\|m_{t,k}\|}{8L} \cdot b_{t,k+1}\right)$ for $\lambda \in [0,1]$. Note that $\gamma'(\lambda) = x_{t,k} - x_t + \frac{\delta\|m_{t,k}\|}{8L} \cdot b_{t,k+1}$. Let $x'_{t,k} = x_{t,k} + \frac{\delta\|m_{t,k}\|}{8L} \cdot b_{t,k+1}$. Since $y_{t,k+1}$ is uniformly sampled from the line segment $\left[x_t, x'_{t,k}\right]$ and f is differentiable at $y_{t,k+1}$ almost surely by Lemma 3.2, we know that

$$\mathbb{E}\left[\langle g_{t,k+1}, x'_{t,k} - x_t\rangle \big| \widehat{\mathcal{F}}_{t,k}\right] = \int_0^1 f'(\gamma(t); x'_{t,k} - x_t)\mathrm{d}t = f(x'_{t,k}) - f(x_t).$$

By $x'_{t,k} - x_t = -\left(1 - \frac{\|m_{t,k}\|}{8L}\right) \cdot \delta \frac{m_{t,k}}{\|m_{t,k}\|} + \frac{\delta\|m_{t,k}\|}{8L} \cdot b_{t,k+1}$, we have

$$\begin{aligned}
&\mathbb{E}\left[\langle g_{t,k+1}, m_{t,k}\rangle \big| \widehat{\mathcal{F}}_{t,k}\right] \\
&= -\frac{\|m_{t,k}\|}{\left(1 - \frac{\|m_{t,k}\|}{8L}\right) \cdot \delta} \cdot \mathbb{E}\left[\langle g_{t,k+1}, x'_{t,k} - x_t\rangle \big| \widehat{\mathcal{F}}_{t,k}\right] + \frac{\|m_{t,k}\|}{\left(1 - \frac{\|m_{t,k}\|}{8L}\right) \cdot \delta} \cdot \mathbb{E}\left[\left\langle g_{t,k+1}, \frac{\delta\|m_{t,k}\|}{8L} \cdot b_{t,k+1}\right\rangle \Big| \widehat{\mathcal{F}}_{t,k}\right] \\
&\le -\frac{\|m_{t,k}\|}{\left(1 - \frac{\|m_{t,k}\|}{8L}\right) \cdot \delta} \cdot \Big(f(x_{t,k}) - f(x_t) - |f(x'_{t,k}) - f(x_{t,k})|\Big) + \frac{\|m_{t,k}\|^2}{8\left(1 - \frac{\|m_{t,k}\|}{8L}\right)} \\
&\le -\frac{\|m_{t,k}\|}{\left(1 - \frac{\|m_{t,k}\|}{8L}\right) \cdot \delta} \cdot \Big(f(x_{t,k}) - f(x_t)\Big) + \frac{\|m_{t,k}\|^2}{4\left(1 - \frac{\|m_{t,k}\|}{8L}\right)},
\end{aligned}$$

which directly implies

$$\mathbb{E}\left[\langle g_{t,k+1}, m_{t,k}\rangle \big| \mathcal{F}_{t,k}\right] \le -\frac{\|m_{t,k}\|}{\left(1 - \frac{\|m_{t,k}\|}{8L}\right) \cdot \delta} \cdot \Big(f(x_{t,k}) - f(x_t)\Big) + \frac{\|m_{t,k}\|^2}{4\left(1 - \frac{\|m_{t,k}\|}{8L}\right)}.$$

By construction, $m_{t,k+1} = \beta m_{t,k} + (1-\beta)g_{t,k+1}$ under $D_{t,k} \cap \cdots \cap D_{t,1}$, and $m_{t,k+1} = 0$ otherwise. Let $\mathcal{D}_{t,k} = D_{t,k} \cap \cdots \cap D_{t,1}$. Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\|m_{t,k+1}\|^2 \big| \mathcal{F}_{t,k}\right] \\
&\le \Big(\beta^2\|m_{t,k}\|^2 + (1-\beta)^2L^2 + 2\beta(1-\beta) \cdot \mathbb{E}\left[\langle g_{t,k+1}, m_{t,k}\rangle \big| \mathcal{F}_{t,k}\right]\Big)\mathbb{1}_{\mathcal{D}_{t,k}} \\
&\le \left(\beta^2\|m_{t,k}\|^2 + (1-\beta)^2L^2 + 2\beta(1-\beta) \cdot \left(-\frac{\|m_{t,k}\|}{\left(1 - \frac{\|m_{t,k}\|}{8L}\right) \cdot \delta} \cdot \Big(f(x_{t,k}) - f(x_t)\Big) + \frac{\|m_{t,k}\|^2}{4\left(1 - \frac{\|m_{t,k}\|}{8L}\right)}\right)\right)\mathbb{1}_{\mathcal{D}_{t,k}} \\
&\le \left(\beta^2\|m_{t,k}\|^2 + (1-\beta)^2L^2 + 2\beta(1-\beta) \cdot \frac{\|m_{t,k}\|^2}{2\left(1 - \frac{\|m_{t,k}\|}{8L}\right)}\right)\mathbb{1}_{\mathcal{D}_{t,k}} =: h(\beta)\mathbb{1}_{\mathcal{D}_{t,k}}.
\end{aligned}$$

By rearranging, it holds

$$h(\beta) = \beta^2 \underbrace{\left(L^2 + \|m_{t,k}\|^2 - \frac{\|m_{t,k}\|^2}{1 - \frac{\|m_{t,k}\|}{8L}} \right)}_{T_1} + \beta \underbrace{\left(\frac{\|m_{t,k}\|^2}{1 - \frac{\|m_{t,k}\|}{8L}} - 2L^2 \right)}_{T_2} + L^2.$$

Note that, by $0 < \|m_{t,k}\| \leq L$ in $\mathcal{D}_{t,k}$ and conditioning on $\mathcal{D}_{t,k}$, it holds

$$T_1 = \frac{L^2}{8\left(1 - \frac{\|m_{t,k}\|}{8L}\right)} \left(8 - \frac{\|m_{t,k}\|}{L} - \frac{\|m_{t,k}\|^3}{L^3} \right) > 0, \quad \text{and} \quad T_2 = \frac{L^2}{4\left(1 - \frac{\|m_{t,k}\|}{8L}\right)} \left(-8 + \frac{\|m_{t,k}\|}{L} + \frac{4\|m_{t,k}\|^2}{L^2} \right) < 0.$$

Thus, $h(\beta)$ achieves the minimum at $\beta_{t,k} = \frac{8L^3 - L^2\|m_{t,k}\| - 4L\|m_{t,k}\|^2}{8L^3 - L^2\|m_{t,k}\| - \|m_{t,k}\|^3}$, which belongs to $\mathcal{F}_{t,k}$. Since $0 < \|m_{t,k}\| \leq L$ in $\mathcal{D}_{t,k}$, we have

$$h(\beta_{t,k}) = \left(1 - c_1 \cdot \frac{\|m_{t,k}\|^2}{L^2}\right) \|m_{t,k}\|^2 \overset{(\sharp)}{\leq} \left(1 - \frac{\|m_{t,k}\|^2}{5L^2}\right) \|m_{t,k}\|^2,$$

where

$$c_1 = \frac{L^2}{\|m_{t,k}\|^2} - \frac{L^2 h(\beta_{t,k})}{\|m_{t,k}\|^4} = \frac{16L^4 - 8L^3\|m_{t,k}\| + L^2\|m_{t,k}\|^2}{64L^4 - 16L^3\|m_{t,k}\| + L^2\|m_{t,k}\|^2 - 8L\|m_{t,k}\|^3 + \|m_{t,k}\|^4}.$$

Let $0 < t := \frac{\|m_{t,k}\|}{L} \leq 1$. For the validity of inequality $(\sharp)$, we observe

$$\frac{1}{5} < \inf_{0<t\leq 1} \frac{16 - 8t + t^2}{64 - 16t + t^2 - 8t^3 + t^4} \leq c_1.$$

To see it, note that, for $0 < t \leq 1$, it holds

$$5 \cdot (16 - 8t + t^2) - (64 - 16t + t^2 - 8t^3 + t^4) = (t+2)\left(10\left(t - \frac{4}{5}\right)^2 + \frac{8}{5} - t^3\right) > 0.$$

Therefore,

$$\mathbb{E}\left[\|m_{t,k+1}\|^2\right] = \mathbb{E}\Big[\mathbb{E}\left[\|m_{t,k+1}\|^2 \middle| \mathcal{F}_{t,k}\right]\Big] \leq \mathbb{E}\Big[\left(1 - \frac{\|m_{t,k}\|^2}{5L^2}\right)\|m_{t,k}\|^2\Big] \leq \left(1 - \frac{\mathbb{E}[\|m_{t,k}\|^2]}{5L^2}\right)\mathbb{E}[\|m_{t,k}\|^2].$$

Then, by a similar argument in the proof of (Zhang et al., 2020, Lemma 13) we have $\mathbb{E}[\|m_{t,K}\|^2] \leq \frac{5L^2}{K+4}$. When $K \geq \frac{80L^2}{\epsilon^2}$, we have $\mathbb{E}[\|m_{t,K}\|^2] \leq \frac{\epsilon^2}{16}$. □

Theorem 3.4. *Let f be L-Lipschitz continuous. Then, Algorithm 1 with $K = \frac{80L^2}{\epsilon^2}$ and $T = \frac{4\Delta}{\epsilon\delta}$ finds an (ϵ, δ)-GAS point with probability $1 - \gamma$ using at most*

$$\frac{320\Delta L^2}{\epsilon^3\delta} \log\left(\frac{4\Delta}{\gamma\epsilon\delta}\right) \qquad \text{oracle calls}$$

with $\mathbb{P}(\texttt{error} = 1) = 0$, where $f(x_0) - \inf_x f(x) \leq \Delta$.

Proof. Using Lemma 3.2 and Lemma 3.3, the remaining parts directly follow from the proof of (Zhang et al., 2020, Theorem 8). □

C. Proofs of Section 3.3

Theorem 3.5. *Under Assumption 3.1(b), with probability at least $\frac{3}{5}$, the output of Algorithm 2 satisfies* $\mathrm{dist}(0, \partial_\delta f(x_{\text{out}})) \leq \epsilon$ *after at most*

$$\widetilde{O}\left(\frac{G^3\Delta}{\epsilon^4\delta}\right) \qquad \text{oracle calls}$$

with $\mathbb{P}(\texttt{error} = 1) = 0$, where $f(x_0) - \inf_x f(x) \leq \Delta$.

Proof. Let $\alpha_i \coloneqq \beta^{t-i}(1-\beta)$ and denote $x'_{t+1} \coloneqq x_{t+1} + \zeta b_{t+1}$ and $\mathcal{G}_t \coloneqq \sigma(g_1, \ldots, g_t), \forall t$. Clearly, the random variables $m_t, x_t, x_{t+1}, \eta_t$ are $\mathcal{G}_t$-measurable. Note that

$$m_t = \beta^K m_{t-K} + \sum_{i=t-K+1}^{t} \alpha_i g_i.$$

Conceptually, if we choose K to be sufficiently large, the term $\beta^K m_{t-K}$ is negligible. Then, if all the points $y_{t-K+1}, \ldots, y_t$ are inside $x_{t-K} + \delta\mathbb{B}$, we have that m_t approximately belongs to $\partial_\delta f(x_{t-K})$ in expectation.

Note that for all $i = t-K+1, \ldots, t$,

$$\begin{aligned}
\|y_i - x_{t-K}\| &\leq \|y_i - x_{i-1}\| + \|x_{i-1} - x_{t-K}\| \\
&\overset{(a)}{\leq} \|x'_i - x_{i-1}\| + \|x_{i-1} - x_{t-K}\| \\
&= \|\zeta b_i - \eta_{i-1} m_{i-1}\| + \left\| \sum_{j=t-K}^{i-2} \eta_j m_j \right\| \\
&\overset{(b)}{\leq} \zeta + \sum_{j=t-K}^{i-1} \eta_j \|m_j\| \\
&\overset{(c)}{\leq} \frac{\omega}{p} + \frac{i-t+K}{p} \\
&\leq \frac{K+\omega}{p},
\end{aligned}$$

where (a) holds since y_i is sampled from the line segment $[x_{i-1}, x'_i]$, (b) uses $\|b_i\| \leq 1$ and (c) follows from $\zeta \leq \frac{\omega}{p}$ and $\eta_t \|m_t\| \leq \frac{1}{p}, \forall t$. We verify that the choices of K, ω and p satisfy $\frac{K+\omega}{p} \leq \delta$:

$$\frac{K+\omega}{p} = \frac{\frac{1}{1-\beta} \ln \frac{16G}{\epsilon}}{\frac{64G^2}{\delta\epsilon^2} \ln \frac{16G}{\epsilon}} = \delta.$$

Then, conditioned on $\mathcal{G}_{t-K}$, since for all $i = t-K+1, \ldots, t$,

$$\mathbb{E}\left[g_i \mid \mathcal{G}_{t-K}\right] = \mathbb{E}\left[\nabla f(y_i) \mid \mathcal{G}_{t-K}\right] \in \partial_\delta f(x_{t-K}),$$

we have (note that $\sum_{i=t-K+1}^{t} \alpha_i = 1 - \beta^K$)

$$\begin{aligned}
&\frac{1}{1-\beta^K} \sum_{i=t-K+1}^{t} \alpha_i \mathbb{E}\left[g_i \mid \mathcal{G}_{t-K}\right] \in \partial_\delta f(x_{t-K}) \\
\Rightarrow\ &\frac{1}{1-\beta^K} \left(\mathbb{E}\left[m_t \mid \mathcal{G}_{t-K}\right] - \beta^K m_{t-K}\right) \in \partial_\delta f(x_{t-K}) \\
\Rightarrow\ &\mathrm{dist}(0, \partial_\delta f(x_{t-K})) \leq \frac{1}{1-\beta^K} \left(\left\|\mathbb{E}\left[m_t \mid \mathcal{G}_{t-K}\right]\right\| + \beta^K \|m_{t-K}\|\right) \\
&\qquad\qquad\qquad\qquad\quad \leq \frac{1}{1-\beta^K} \left(\mathbb{E}\left[\|m_t\| \mid \mathcal{G}_{t-K}\right] + \beta^K \|m_{t-K}\|\right).
\end{aligned}$$

Take expectation.

$$\begin{aligned}
\mathbb{E}\left[\mathrm{dist}(0, \partial_\delta f(x_{t-K}))\right] &\leq \frac{1}{1-\beta^K} \mathbb{E}\left[\|m_t\|\right] + \frac{\beta^K G}{1-\beta^K}, \\
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[\mathrm{dist}(0, \partial_\delta f(x_{t-K}))\right] &\leq \frac{1}{(1-\beta^K)T} \sum_{t=1}^{T} \mathbb{E}\left[\|m_t\|\right] + \frac{\beta^K G}{1-\beta^K}.
\end{aligned}$$

We verify that the choices of β and K satisfy $\beta^K G \le \frac{\epsilon}{16}$: $\left(\beta^K \le \frac{\epsilon}{16G}\right) \Leftrightarrow \left(K \ge \frac{1}{\ln\frac{1}{\beta}} \ln \frac{16G}{\epsilon}\right)$. WLOG, we assume that $\epsilon \le G$, and thus $\beta^K \le \frac{1}{16}$. The above inequality can be further bounded as

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\operatorname{dist}(0, \partial_\delta f(x_{t-K}))\right] \le \frac{16}{15T}\sum_{t=1}^{T}\mathbb{E}\left[\|m_t\|\right] + \frac{\epsilon}{15}. \tag{C.1}$$

The remaining proof is to show that Algorithm 2 ensures that $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|m_t\|\right] = O(\epsilon)$.

For ease of analysis, we denote $\mathcal{Y}_{t+1} := \sigma(g_1, \ldots, g_t, b_{t+1}, y_{t+1})$ and $\hat{\mathcal{Y}}_{t+1} := \sigma(g_1, \ldots, g_t, b_{t+1})$. Clearly, we have $\mathcal{G}_t \subset \hat{\mathcal{Y}}_{t+1} \subset \mathcal{Y}_{t+1} \subset \mathcal{G}_{t+1}$. Let $\varphi(\lambda) := (1-\lambda)x_t + \lambda x'_{t+1}$ for $\lambda \in [0,1]$. Since y_{t+1} is uniformly sampled from the line segment $[x_t, x'_{t+1}]$ and that f is differentiable at y_{t+1} almost surely, it holds that

$$\begin{aligned}
\mathbb{E}\left[\left\langle g_{t+1}, x'_{t+1} - x_t\right\rangle \mid \mathcal{G}_t\right] &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[\left\langle g_{t+1}, x'_{t+1} - x_t\right\rangle \mid \mathcal{Y}_{t+1}\right] \mid \hat{\mathcal{Y}}_{t+1}\right] \mid \mathcal{G}_t\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left\langle \nabla f(y_{t+1}), x'_{t+1} - x_t\right\rangle \mid \hat{\mathcal{Y}}_{t+1}\right] \mid \mathcal{G}_t\right] \\
&= \mathbb{E}\left[\int_0^1 f'(\varphi(\lambda); x'_{t+1} - x_t)\mathrm{d}\lambda \mid \mathcal{G}_t\right] \\
&= \mathbb{E}\left[f(x'_{t+1}) - f(x_t) \mid \mathcal{G}_t\right].
\end{aligned} \tag{C.2}$$

By $x'_{t+1} - x_t = -\eta_t m_t + \zeta b_{t+1}$, we have

$$\begin{aligned}
\mathbb{E}\left[\left\langle g_{t+1}, x'_{t+1} - x_t\right\rangle \mid \mathcal{G}_t\right] &= -\eta_t \mathbb{E}\left[\langle g_{t+1}, m_t\rangle \mid \mathcal{G}_t\right] + \zeta\mathbb{E}\left[\langle g_{t+1}, b_{t+1}\rangle \mid \mathcal{G}_t\right] \\
&\le -\eta_t \mathbb{E}\left[\langle g_{t+1}, m_t\rangle \mid \mathcal{G}_t\right] + \zeta G,
\end{aligned}$$

where we used $\|b_{t+1}\| \le 1$. Thus, combining with (C.2), we obtain

$$\begin{aligned}
\mathbb{E}\left[\langle g_{t+1}, m_t\rangle \mid \mathcal{G}_t\right] &\le \frac{1}{\eta_t}\mathbb{E}\left[f(x_t) - f(x_{t+1}) + f(x_{t+1}) - f(x'_{t+1}) \mid \mathcal{G}_t\right] + \frac{\zeta}{\eta_t}G \\
&\le \frac{1}{\eta_t}\big(f(x_t) - f(x_{t+1})\big) + \frac{\zeta}{\eta_t}(L+G).
\end{aligned} \tag{C.3}$$

Based on the construction $m_{t+1} = \beta m_t + (1-\beta)g_{t+1}$, we can conclude that

$$\begin{aligned}
\|m_{t+1}\|^2 &= \beta^2\|m_t\|^2 + 2\beta(1-\beta)\langle g_{t+1}, m_t\rangle + (1-\beta)^2\|g_{t+1}\|^2, \\
\mathbb{E}\left[\eta_t\big(\|m_{t+1}\|^2 - \beta^2\|m_t\|^2\big)\right] &= 2\beta(1-\beta)\mathbb{E}\left[\eta_t\langle g_{t+1}, m_t\rangle\right] + (1-\beta)^2\mathbb{E}\left[\eta_t\|g_{t+1}\|^2\right].
\end{aligned}$$

From (C.3), it holds that

$$\begin{aligned}
\mathbb{E}\left[\eta_t\big(\|m_{t+1}\|^2 - \beta^2\|m_t\|^2\big)\right] &\le 2\beta(1-\beta)\mathbb{E}\left[f(x_t) - f(x_{t+1})\right] + 2\beta(1-\beta)(L+G)\zeta \\
&\quad + (1-\beta)^2\mathbb{E}\left[\eta_t\|g_{t+1}\|^2\right], \\
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\eta_t\big(\|m_{t+1}\|^2 - \beta^2\|m_t\|^2\big)\right] &\le \frac{2\beta(1-\beta)\Delta}{T} + 2\beta(1-\beta)(L+G)\zeta + \frac{(1-\beta)^2G^2}{q},
\end{aligned}$$

where we used $\eta_t \le \frac{1}{q}$.

Since $\eta_t = \frac{1}{p\|m_t\|+q}$, using the same telescoping proof in (Zhang et al., 2020), as long as $\frac{pG}{q} \le \frac{\beta}{2}$, the following holds

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\eta_t\big(\|m_{t+1}\|^2 - \beta^2\|m_t\|^2\big)\right] \ge \frac{\beta(1-\beta)}{2T}\sum_{t=1}^{T+1}\mathbb{E}\left[\frac{\|m_t\|^2}{p\|m_t\|+q}\right] - \frac{\beta G^2}{qT}.$$

Thus,

$$\frac{\beta(1-\beta)}{2T}\sum_{t=1}^{T+1}\mathbb{E}\left[\frac{\|m_t\|^2}{p\|m_t\|+q}\right]\leq\frac{2\beta(1-\beta)\Delta}{T}+2\beta(1-\beta)(L+G)\zeta+\frac{(1-\beta)^2G^2}{q}+\frac{\beta G^2}{qT},$$
$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\frac{q\|m_t\|^2}{p\|m_t\|+q}\right]\leq\frac{4q\Delta}{T}+4q(L+G)\zeta+\frac{2(1-\beta)G^2}{\beta}+\frac{2G^2}{T(1-\beta)}.$$

Comparing the above inequality with (14)[5] in (Zhang et al., 2020), we notice that the only difference is the additional perturbation term $4q(L+G)\zeta$. Since we choose the identical β, p, q and T as in (Zhang et al., 2020), using the arguments (15) and (16) in (Zhang et al., 2020) and denoting $m_{avg} := \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|m_t\|\right]$, we obtain

$$\frac{4Gm_{avg}^2}{m_{avg}+4G}\leq\frac{\epsilon^2}{17}+4q(L+G)\zeta \overset{(\star)}{\leq}\frac{\epsilon^2}{15},$$

where $(\star)$ uses $\zeta\leq\frac{\epsilon^2}{510q(L+G)}$. The above is a quadratic equation in m_{avg}:

$$4Gm_{avg}^2-\frac{\epsilon^2}{15}m_{avg}-\frac{4G\epsilon^2}{15}\leq 0.$$

Solving for the positive root of this quadratic equation and using $\epsilon\leq G$, we obtain

$$m_{avg}\leq\frac{\frac{\epsilon^2}{15}+\sqrt{\frac{\epsilon^4}{225}+\frac{64G^2\epsilon^2}{15}}}{8G}\leq\frac{4\epsilon}{15}\leq\frac{5\epsilon}{16}.$$

Finally, using (C.1), we conclude that

$$\mathbb{E}\left[\text{dist}(0,\partial_\delta f(x_{out}))\right]=\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\text{dist}(0,\partial_\delta f(x_{t-K}))\right]\leq\frac{2\epsilon}{5}.$$

Thus, with probability at least $\frac{3}{5}$, we have $\text{dist}(0,\partial_\delta f(x_{out}))\leq\epsilon$. □

D. Proofs of Section 4.1

Theorem 4.4 (NAS by GAS)**.** *For a Lipschitz continuous f, suppose that $\partial f:\mathbb{R}^d\rightrightarrows\mathbb{R}^d$ is (δ,η,κ)-OLC. If x is (ϵ,η)-GAS, then x is $\big(\epsilon+\kappa(\delta+\eta),\delta\big)$-NAS.*

Proof. As x is Goldstein (ϵ,η)-stationary, we have $\text{dist}\big(0,\partial f(x+\eta\mathbb{B})\big)\leq\epsilon$, which implies that there exists

$$\|g\|\leq\epsilon,\qquad\text{such that}\qquad g\in\partial f(x+\eta\mathbb{B})=\text{Co}\left\{\bigcup_{y\in\mathbb{B}_\eta(x)}\partial f(y)\right\}.$$

By Carathéodory's theorem (Rockafellar & Wets, 2009, Theorem 2.29), we can write $g=\sum_{j=1}^{d+1}\alpha_j g_j$, where $\alpha_j\geq 0,\sum_{j=1}^{d+1}\alpha_j=1, g_j\in\partial f(y_j), y_j\in\mathbb{B}_\eta(x),\forall j\in[d+1]$.

Let $y\in\mathbb{B}_\delta(x)$ be a pivot such that ∂f is κ-outer Lipschitz continuous on $\mathbb{B}_\eta(x)$. As f is Lipschitz and by (Clarke, 1990, Proposition 2.1.2), $\partial f(y)$ is nonempty, convex, and compact. Let $u_j:=\arg\min_{z\in\partial f(y)}\|z-g_j\|, u:=\sum_{j=1}^{d+1}\alpha_j u_j\in\partial f(y)$. Then, we compute

$$\|u\|=\left\|\sum_{j=1}^{d+1}\alpha_j u_j\right\|\leq\|g\|+\sum_{j=1}^{d+1}\alpha_j\|u_j-g_j\|\leq\|g\|+\kappa\sum_{j=1}^{d+1}\alpha_j\|y-y_j\|\leq\|g\|+\kappa\sum_{j=1}^{d+1}\alpha_j\big(\|y-x\|+\|x-y_j\|\big)\leq\epsilon+\kappa(\delta+\eta),$$

[5]There is a typo in the telescoping proof of Theorem 14 in (Zhang et al., 2020): The term $\frac{\beta^2G^2}{q}$ above Equation (14) should be $\frac{\beta G^2}{q}$. This typo does not affect the final convergence result.

which completes the proof. ☐

Theorem 4.5. *Let $\delta > 0$ and $\partial f : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be κ-OLC. For any compact set S, there exists an $\eta \in (0, \delta]$ such that ∂f is (δ, η, κ)-OLC on S.*

Proof. Let $\bigcup_{x \in S} \mathbb{B}^\circ_{\frac{1}{2}\bar{\delta}(x)}(x)$ be an open cover of S, where $\bar{\delta}(x) = \min\{\delta, \delta(x)\}$ and $\delta(x)$ is the inradius of neighborhood $V(x)$, on which ∂f is κ-outer Lipschitz at x, satisfying $\mathbb{B}_{\delta(x)}(x) \subseteq V(x)$. As S is compact, we find a finite subcover $\bigcup_{i \in [m]} \mathbb{B}^\circ_{\frac{1}{2}\delta_i}(x_i)$, where $\delta_i = \min\{\delta, \delta(x_i)\}$ and $x_i \in S$. Let $\eta := \min_{i \in [m]} \frac{\delta_i}{2}$. Then, by Lebesgue's number theorem (Munkres, 1974, Chapter 3, Lemma 7.2) on open cover $\bigcup_{x \in S} \mathbb{B}^\circ_{\bar{\delta}(x)}(x)$ of S, for any $x \in S$, there exists $i \in [m]$ such that $\mathbb{B}^\circ_\eta(x) \subseteq \mathbb{B}^\circ_{\delta_i}(x_i)$. Thus, $\mathbb{B}_\eta(x) \subseteq \mathbb{B}_{\delta_i}(x_i) \subseteq V(x_i)$. For any $z \in \mathbb{B}_\eta(x) \cap S$, by κ-outer Lipschitz continuity ∂f on $V(x_i)$, we have

$$\partial f(z) \subseteq \partial f(x_i) + \kappa \|x_i - z\| \mathbb{B}, \; \forall z \in \mathbb{B}_\eta(x) \cap S,$$

where $x_i \in \mathbb{B}_{\delta_i}(x) \subseteq \mathbb{B}_\delta(x)$. This completes the proof. ☐

E. Proofs of Section 4.2

Proposition 4.7 (smooth regularization)**.** *Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ has a (δ, η, κ)-OLC ∂f and $g : \mathbb{R}^d \to \mathbb{R}$ is differentiable with a β-Lipschitz gradient ∇g. Then $\partial(f + g)$ is $(\delta, \eta, \beta + \kappa)$-OLC.*

Proof. Let $F := f + g$. By (Rockafellar & Wets, 2009, Exercise 8.8(c)), $\partial F = \partial f + \nabla g$. Let $y \in \mathbb{B}_\delta(x)$ be a pivot of ∂f. Then, for $\forall z \in \mathbb{B}_\eta(x)$, we compute

$$\partial F(z) = \partial f(z) + \nabla g(z) \subseteq \partial f(z) + \nabla g(y) + \beta \|z - y\| \mathbb{B} \subseteq \partial F(y) + (\beta + \kappa)\|z - y\| \mathbb{B},$$

which completes the proof. ☐

Proposition 4.8 (separable sum)**.** *Suppose, for any $i \in [m]$, that $f_i : \mathbb{R}^{d_i} \to \mathbb{R}$ has a $(\delta_i, \eta_i, \kappa_i)$-OLC ∂f_i. Let $f(x) := \sum_{i=1}^m f_i(x_i)$, where $x := \bigoplus_{i=1}^m x_i$. Then, ∂f is (δ, η, κ)-OLC with*

$$\delta = \sqrt{\sum_{i=1}^m \delta_i^2}, \quad \eta = \min_{i \in [m]} \eta_i, \quad \kappa = \sqrt{\sum_{i=1}^m \kappa_i^2}.$$

Proof. By (Rockafellar & Wets, 2009, Proposition 10.5) and f is Lipschitz, $\partial f = \bigoplus_{i=1}^m \partial f_i$. Let $y_i \in \mathbb{B}^{d_i}_{\delta_i}(x_i)$ be a pivot of ∂f_i. Also $y := \bigoplus_{i=1}^m y_i$. Similarly, for any $z \in \mathbb{B}_\eta(x)$, it holds $z_i \in \mathbb{B}^{d_i}_\eta(x_i) \subseteq \mathbb{B}^{d_i}_{\eta_i}(x_i), \forall i \in [m]$. We compute

$$\partial f(z) = \bigoplus_{i=1}^m \partial f_i(z_i) \subseteq \bigoplus_{i=1}^m \left(\partial f_i(y_i) + \kappa_i \|y_i - z_i\| \mathbb{B}^{d_i} \right) \subseteq \partial f(y) + \kappa \|z - y\| \mathbb{B}^d,$$

where $\|y - x\|^2 = \sum_{i=1}^m \|y_i - x_i\|^2 \le \sum_{i=1}^m \delta_i^2 = \delta^2$. This completes the proof. ☐

Proposition 4.9 (linear composition)**.** *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ has a (δ, η, κ)-OLC ∂f and $A \in \mathbb{R}^{n \times d}$ is surjective. Then, $\partial(f \circ A)$ is $\left(\delta \|A^\dagger\|, \frac{\eta}{\|A\|}, \kappa \|A\|^2\right)$-outer Lipschitz continuous.*

Proof. Let $F(x) := f(Ax)$. As A is surjective, by (Rockafellar & Wets, 2009, Exercise 10.7), $\partial F(x) = A^\top \partial f(Ax)$. Let $q \in \mathbb{B}^n_\delta(Ax)$ be a pivot of ∂f. Let $y := A^\dagger q + (I - A^\dagger A)x$. Then $Ay = q$ and $\|y - x\| \le \|A^\dagger\| \|q - Ax\| \le \delta \|A^\dagger\|$. Meanwhile, for any $z \in \mathbb{B}^d_{\frac{\eta}{\|A\|}}(x)$, it holds $\|Az - Ax\| \le \|A\| \|z - x\| \le \eta$. We compute

$$\partial F(z) = A^\top \partial f(Az) \subseteq A^\top \partial f(Ay) + \kappa \|Ay - Az\| A^\top \mathbb{B}^n \subseteq \partial F(y) + \kappa \|A\|^2 \|y - z\| \mathbb{B}^d,$$

which completes the proof. ☐

Proposition 4.10 (rescaling)**.** *Suppose that the L_1-Lipschitz $f : \mathbb{R}^n \to \mathbb{R}$ has a (δ, η, κ)-OLC ∂f and $g : \mathbb{R} \to \mathbb{R}$ is L_2-Lipschitz and β-smooth. Then, $\partial(g \circ f)$ is $(\delta, \eta, \beta L_1 + \kappa L_2)$-OLC.*

Proof. Let $F = g \circ f$. By (Clarke, 1990, Theorem 2.3.9(ii)), $\partial F(x) = \nabla g(f(x)) \cdot \partial f(x)$. Let $y \in \mathbb{B}_\delta(x)$ be a pivot of ∂f. Then, for $\forall z \in \mathbb{B}_\eta(x)$, we compute

$$\begin{aligned}\partial F(z) = \nabla g(f(z)) \cdot \partial f(z) &\subseteq \nabla g(f(z)) \cdot \partial f(y) + \nabla g(f(z)) \cdot \kappa\|z-y\|\mathbb{B} \\ &\subseteq \nabla g(f(y)) \cdot \partial f(y) + (\beta L_1 + \kappa L_2)\|z-y\|\mathbb{B} \\ &= \partial F(y) + (\beta L_1 + \kappa L_2)\|z-y\|\mathbb{B},\end{aligned}$$

which completes the proof. □

Proposition 4.11 (sum)**.** *Suppose, for any $i \in [m]$, that $G_i : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is $(\delta_i, \eta_i, \kappa_i)$-OLC with a shared pivot mapping $P : \mathbb{R}^d \to \mathbb{R}^d$. Let $G(x) \coloneqq \sum_{i=1}^m G_i(x)$. Then, G is (δ, η, κ)-OLC with*

$$\delta = \min_{i\in[m]} \delta_i, \quad \eta = \min_{i\in[m]} \eta_i, \quad \kappa = \sum_{i=1}^m \kappa_i.$$

Proof. Let $y = P(x) \in \mathbb{B}_{\delta_i}(x)$ be a pivot of $G_i(x)$, which by pivot sharing assumption should hold for all $i \in [m]$. Thus $\|y - x\| \le \min_{i\in[m]} \delta_i = \delta$. For all $z \in \mathbb{B}_\eta(x) \subseteq \mathbb{B}_{\eta_i}(x)$, we compute

$$G(z) = \sum_{i=1}^m G_i(z) \subseteq \sum_{i=1}^m \Big(G_1(y) + \kappa_i\|z-y\|\mathbb{B}\Big) \subseteq G(y) + \left(\sum_{i=1}^m \kappa_i\right)\|z-y\|\mathbb{B},$$

as expected. □

Corollary 4.12 (partially separable sum)**.** *Suppose, for any $i \in [m]$, that $G_i : \mathbb{R}^{d_0} \times \mathbb{R}^{d_i} \rightrightarrows \mathbb{R}^{d_0} \times \mathbb{R}^{d_i}$ is $(\delta_i, \eta_i, \kappa_i)$-OLC with a partially shared pivot mapping $P_i : \mathbb{R}^{d_0} \times \mathbb{R}^{d_i} \to \mathbb{R}^{d_0} \times \mathbb{R}^{d_i}$, such that $\pi_1 \circ P_i(x_0, x_i) = \pi_1 \circ P_1(x_0, x_1), \forall i \in [m]$. Let $G(x) \coloneqq \sum_{i=1}^m G_i(x_0, x_i)$, where $x \coloneqq \bigoplus_{i=0}^m x_i$. Then, G is (δ, η, κ)-OLC with*

$$\delta = \sqrt{\sum_{i=1}^m \delta_i^2}, \quad \eta = \min_{i\in[m]} \eta_i, \quad \kappa = \sum_{i=1}^m \kappa_i.$$

Proof. Let $(y_0, y_i) \in \mathbb{B}_{\delta_i}^{d_0+d_i}\big((x_0, x_i)\big)$ be a pivot of G_i. Also $y \coloneqq \bigoplus_{i=0}^m y_i$. Similarly, for any $z \in \mathbb{B}_\eta(x)$, it holds $(z_0, z_i) \in \mathbb{B}_\eta^{d_0+d_i}\big((x_0, x_i)\big) \subseteq \mathbb{B}_{\eta_i}^{d_0+d_i}\big((x_0, x_i)\big), \forall i \in [m]$. We compute

$$G(z) = \sum_{i=1}^m G_i\big((z_0, z_i)\big) \subseteq \sum_{i=1}^m \Big(G_i\big((y_0, y_i)\big) + \kappa_i\big\|(z_0, z_i) - (y_0, y_i)\big\|\mathbb{B}^{d_0+d_i}\Big) \subseteq G(y) + \kappa\|z-y\|\mathbb{B}^d,$$

where $\|y - x\|^2 = \sum_{i=0}^m \|y_i - x_i\|^2 \le \sum_{i=1}^m \big(\|y_0 - x_0\|^2 + \|y_i - x_i\|^2\big) \le \sum_{i=1}^m \delta_i^2 = \delta^2$, and $d = \sum_{i=0}^m d_i$. This completes the proof. □

F. Proofs of Section 5

Claim 5.1. *For $\varrho(u_1, u_2) \coloneqq u_1 \cdot \max\{u_2, 0\}$, it holds*

$$\partial\varrho(u_1, u_2) = \begin{cases} (u_2, u_1) & \text{for} \quad u_2 > 0, \\ (0, 0) & \text{for} \quad u_2 < 0, \\ (0, \mathrm{Co}\{0, u_1\}) & \text{for} \quad u_2 = 0. \end{cases}$$

Proof. Define

$$\begin{aligned} C_1 &\coloneqq \{(u_1, u_2) : u_2 \ge 0\}, \\ C_2 &\coloneqq \{(u_1, u_2) : u_2 \le 0\}. \end{aligned}$$

It is clear that $C_1 \cup C_2 = \mathbb{R}^2$, and we have

$$\varrho(u_1, u_2) = \begin{cases} u_1 \cdot u_2 & \text{for} \quad (u_1, u_2) \in C_1, \\ 0 & \text{for} \quad (u_1, u_2) \in C_2. \end{cases}$$

Note that $C_1 \cap C_2$ form a set S of measure 0, and if $(u_1, u_2) \notin S$, then ϱ is differentiable. The claim follows from taking convex hull with (Rockafellar & Wets, 2009, Theorem 9.61). □

In the following proof, we will use the following notion named partial Clarke subdifferential. See also (Clarke, 1990, Page 48), (Rockafellar & Wets, 2009, Corollary 10.11).

Definition F.1. *Let a local Lipschitz function $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ and $g_y : x \to f(x, y)$. Then the partial Clarke subdifferential with respect to the first argument is defined as $\partial_1 f(x, y) := \partial g_y(x)$. $\partial_2 f(x, y)$ is defined similarly.*

Claim F.2. $\partial\varrho(u_1, u_2) = \partial_1\varrho(u_1, u_2) \times \partial_2\varrho(u_1, u_2)$ *and* $|\pi_1 \circ \partial\varrho(u_1, u_2)| = 1$.

Proof. Note that $\partial_1\varrho(u_1, u_2) = \max\{u_2, 0\}$ and $\partial_2\varrho(u_1, u_2) = u_1 \cdot \partial(\max\{\cdot, 0\})(u_2)$. The proof completes by using Claim 5.1 and literally checking definitions. □

Proposition 5.2 (partially differentiable sum rule)**.** *It holds*

$$\partial f(a, U) = \sum_{i=1}^{n} \partial h_i(a, u_i).$$

Proof. To begin, we observe the following general fact. For any set $A \subseteq \mathbb{R}^n \times \mathbb{R}^m$, if $|\pi_1 A| = 1$, then $A = \pi_1 A \times \pi_2 A$. To see it, for one direction, if $(a_1, a_2) \in A$, then $a_1 \in \pi_1 A, a_2 \in \pi_2 A$. Thus, $A \subseteq \pi_1 A \times \pi_2 A$. For the other direction, let $a_1 \in \pi_1 A, a_2 \in \pi_2 A$. As $\{a_1\} = \pi_1 A$, then by the definition of $\pi_2 A$, it holds $(a_1, a_2) \in A$. Thus $\pi_1 A \times \pi_2 A \subseteq A$.

To avoid uninformative sophisticated notation, we will use "$\overset{P}{=}$" for equivalence up to coordinate permutation. Formally, if $A \overset{P}{=} B$, then there exists a permutation matrix P such that $B = \{Px : x \in A\}$. We compute

$$\begin{aligned}
\partial f(a, U) &\subseteq \sum_{i=1}^{n} \partial h_i(a, u_i) && \text{(Clarke (1990, Proposition 2.3.3))} \\
&= \sum_{i=1}^{n} \nabla \ell_i \left(\sum_{j=1}^{m} \varrho(a_j, u_{ij}) \right) \cdot \bigoplus_{j=1}^{m} \partial\varrho(a_j, u_{ij}) && \text{(Clarke (1990, Theorem 2.3.9(ii)))} \\
&= \sum_{i=1}^{n} \nabla \ell_i \left(\sum_{j=1}^{m} \varrho(a_j, u_{ij}) \right) \cdot \bigoplus_{j=1}^{m} \Big(\partial_1\varrho(a_j, u_{ij}) \times \partial_2\varrho(a_j, u_{ij}) \Big) && \text{(Claim F.2)} \\
&\overset{P}{=} \underbrace{\left(\sum_{i=1}^{n} \nabla \ell_i \left(\sum_{j=1}^{m} \varrho(a_j, u_{ij}) \right) \cdot \bigoplus_{j=1}^{m} \partial_1\varrho(a_j, u_{ij}) \right)}_{S_1} \times \underbrace{\left(\bigoplus_{i=1}^{n} \bigoplus_{j=1}^{m} \nabla \ell_i \left(\sum_{j=1}^{m} \varrho(a_j, u_{ij}) \right) \cdot \partial_2\varrho(a_j, u_{ij}) \right)}_{S_2}.
\end{aligned}$$

Note that $|S_1| = 1$ as $|\partial_1\varrho(a_j, u_{ij})| = 1, \forall (i, j) \in [n] \times [m]$ by Claim F.2. Thus $|\pi_1 \circ \partial f(a, U)| = 1$ and $\partial f(a, U) \overset{P}{=} \pi_1 \circ \partial f(a, U) \times \pi_2 \circ \partial f(a, U)$. With (Clarke, 1990, Proposition 2.3.16), we compute

$$\begin{aligned}
\partial f(a, U) &\overset{P}{=} \Big(\pi_1 \circ \partial f(a, U) \Big) \times \Big(\pi_2 \circ \partial f(a, U) \Big) \\
&\supseteq \partial_1 f(a, U) \times \partial_2 f(a, U) \\
&\overset{\natural}{=} S_1 \times S_2.
\end{aligned}$$

To see ($\natural$), note that $f(\cdot, U)$ is differentiable. Thus it is straightforward to check $S_1 = \partial_1 f(a, U)$. For S_2, note that $f(a, \cdot)$ is fully separable (as a is fixed). Then, with (Rockafellar & Wets, 2009, Proposition 10.5) and (Clarke, 1990, Theorem 2.3.9(ii)), the verification of $S_2 = \partial_2 f(a, U)$ is routine.

This completes the proof. □